Network-based approach for Post Genome-Wide Association Study Analysis in Admixed Populations

By Mamana Mbiyavanga
(mamana.mbiyavanga@uct.ac.za)
Supervision: Ass. Prof. Nicola Mulder
(nicola.mulder@uct.ac.za)

Computational Biology Group
Institute for Infectous Diseases and Molecular Medicine
University of Cape Town

A thesis submitted for the degree of
Master's in Medicine/Bioinformatics
April 2014

Abstract

The rapid advances in genotyping technology has increased the power to identify loci associated with a complex trait through genome-wide association (GWA) studies. Despite the success of GWA studies that has enabled the identification of associations between common genetic variants and complex diseases, which generally consider only the most significant SNPs/genes, such a single-marker-based approach has shown certain limitations. Therefore it might not possess adequate power to detect important genetic variants, mostly with relatively small effect sizes in complex diseases. As a consequence, the estimated heritability remains unexplained by these variants for most complex diseases covered by published GWAS so far. Extending GWA study findings and exploring related discoveries in a more mechanistic way, taking into account the genetic architecture of the disease as well as the genetic mechanism involved in the disease pathogenesis, to uncover biological pathways and biological networks relevant to phenotypes, should involve combining weak signals from the individual variants from GWA studies to approximate the true disease process more closely and provide biological insights. This would facilitate better understanding of the biological basis of disease susceptibility and using these genetic risk factors to make predictions about who is at risk, in order to develop new prevention and treatment strategies such as new pharmacologic therapies and personalized medicine. Recently, several studies have explored the feasibility of pathway or network-based analysis approaches for GWA studies and they have proposed several methods to summarize the significance of a set of genes or a biological pathway from a collection of SNPs and to adjust for multiple testing at both the gene and pathway levels.

In this project, we review some existing pathway-based approaches for GWA study analyses, by exploring different implemented methods for combining effects of multiple modest genetic variants at gene and pathway levels. We then propose a graph-based method, ancGWAS, that incorporates the signal from GWA study, and the locus-specific ancestry into the human protein-protein interaction (PPI) network to identify significant sub-networks or pathways associated with the trait of interest. This network-based method applies centrality measures within linkage disequilibrium (LD) on the network to search for pathways and applies a scoring

summary statistic on the resulting pathways to identify the most enriched pathways associated with complex diseases. In addition, the proposed method also tests for possible signals of unusual differences in excess/deficiency of ancestry at gene and pathway levels in admixed populations. Through simulations of well-characterized heterogeneous populations, we evaluated and compared the developed method with some existing methods, and demonstrated that this approach may enable more efficient and more powerful analysis of GWA studies, compared to most existing methods, as it incorporates topological properties of networks that provide more information on the relatedness and interconnectivity of genes. We applied the new approach to the GWA study dataset from the Cancer Genetic Markers of Susceptibility (CGEMS) for postmenopausal women of European ancestry with invasive breast cancer. Our analysis on the CGEMS breast cancer data revealed some previously targeted breast cancer pathways, and many others believed to be involved in breast carcinogenesis including, the *Proteaglycan syndecan-1-mediated signaling pathway*, the *ErbB receptor signaling network*, the *Regulation of Androgen receptor activity* pathway, and the *Integrin family cell surface interactions pathway*. The results suggest that genetic alterations in these pathways may contribute to breast cancer susceptibility.

Declaration

This thesis:

- is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specified in the text;
- is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: .....

Date: .....

Mamana Mbiyavanga

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Prof. Nicola Mulder for the continuous support of my master study and research, for her patience, enthusiasm, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a such better supervisor and mentor for my master study.
My sincere thanks also goes to Prof. Darren Martin, Dr. Chimusa Emile, and Dr. Mazandu Gaston, for the support on the useful comments, remarks and engagement through the learning process of this master thesis.

My master studies was supported by a UCT-AIMS joint grant from the African Institute for Mathematical Sciences and the University of Cape Town. Travel grants from the University of Cape Town allowed me to present some of this work at an international conference as well as to attend practical workshops related to my study and research.

I thank my fellow labmates in the Computational Biology Group of the University of Cape Town for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last two years.

I would also like to thank my friends and my loved ones, Nkosi Dimba, Kilandamoko Ephraim, Gracia Nginamau, Matondo Christelle and Luzolawu Prisca, who have supported me throughout entire process, both by keeping me harmonious and helping me putting pieces together. I will be grateful forever for your love. Last but not the least, I would like to thank my family: my parents Nginamau Filipe and Kisuvidi Lando, for giving birth to me at the first place and supporting me spiritually throughout my life.

Contents

List of Figures

List of Tables

Motivation and outline

The rapid advances in genotyping technology have increased the power to identify loci associated with a complex trait through genome-wide association (GWA) studies, which examine, at the genomic level, the relationship between each Single Nucleotide Polymorphism (SNP) marker and the trait of interest. GWAS have been successful during the past few years and through this experimental design, many novel genes have been identified to be significantly and strongly associated with one or more complex trait, leading to important contributions to the understanding of the biological underpinnings of disease susceptibility for a broad spectrum of complex traits (Iles, M. M., 2008; Hindorff, L. A. *et al.*, 2013a). It is now well known that genes do not operate in isolation, rather they collaborate in groups to carry out specific biological functions. Therefore, depending on the genetic architecture of a complex disease, it is possible that many SNPs or genes having low or moderate risk interact to confer a significant combined effect for the complex disease (Schadt, E. E., 2009). Many recent studies have investigated and demonstrated the role of complex molecular networks and cellular pathways in the pathogenesis of complex diseases such as the noninfectious disease cancer (Collins, A. *et al.*, 2011; Filmus, J., 2001; Leivonen, M. *et al.*, 2004), and the autoimmune disease rheumatoid arthritis (RA) (Martinez, A. *et al.*, 2006). For instance, a highly significant interaction was detected between genes associated with the infectious disease tuberculosis (TB) (de Wit, E. *et al.*, 2011). In other examples, genes of the complement pathway are suspected to be involved in the disease pathogenesis of age-related macular degeneration (AMD) and the autophagy pathway was suspected to be implicated in the inflammatory Crohn's disease (CD) pathogenesis (Wang, K. *et al.*, 2010).

Motivation and purpose of the study

Despite the success of GWA studies, which generally consider only the most significant SNPs, at the identification of associations between common genetic variants and complex traits, such a single-marker-based approach has shown certain limitations and might therefore not possess adequate power to detect important genetic variants, with relatively small effect sizes in complex diseases. As a consequence, the disease heritability remains unexplained for almost all the complex diseases covered by published GWAS so far (Iles, M. M., 2008; Wang, K. *et*

al., 2010; Akira, S. *et al.*, 2004; Visscher, P. M., 2008). Extending GWA study findings and exploring related discoveries in a more mechanistic way, by taking into account the genetic architecture of the disease, to uncover biological pathways and biological networks relevant to phenotypes, would involve combining weak signals from the individual variants from GWA studies to approximate the true disease process more closely. Knowing pathways involved in disease pathogenesis may help to tackle disease progression more efficiently. This would facilitate better understanding of the biological basis of disease susceptibility and enable using these genetic risk factors to make predictions about who is at risk, in order to develop new prevention and treatment strategies and move towards personalized medicine (Ramanan, V. K. *et al.*, 2012).

Proposed methodology and study outline

Recently, several studies have explored the feasibility of pathway or network-based analysis approaches for GWA studies and they have proposed several methods to summarize the significance of a set of genes or biological pathway from a collection of SNPs and to adjust for multiple testing at both the gene and pathway levels (Schadt, E. E., 2009; Jia, P. *et al.*, 2011; Wang, K. *et al.*, 2010; Ramanan, V. K. *et al.*, 2012). Pathway-based approaches jointly consider multiple contributing factors in the same pathway and examine whether a group of genes in the same pathway are jointly associated with the trait of interest. Using prior biological knowledge on gene functions to facilitate more powerful analysis of GWAS datasets, these approaches have become increasingly popular and invaluable tools enabling powerful association tests. Thus, they might complement the most-significant SNPs/genes approach to provide additional insights into the interpretation of GWA studies and help to formulate new hypotheses on disease susceptibility and disease progression (Wang, K. *et al.*, 2010). Most existing pathway-based approaches for GWA study data analysis have focused on thoroughly examining a collection of predefined gene sets or pathways based on certain prior biological knowledge, and summarize the significance of each pathway based on the association with the disease of markers in or near genes that are components of the pathway. However, accounting for biological network topology and properties, gene-gene interactions may also enable more powerful analysis of GWA study data sets, compared to analysis of groups of distinct genes, as networks provide more information on the relatedness and interconnectivity of genes (Wang, K. *et al.*, 2010).

Here we review some existing pathway-based approaches for GWA study analyses, by exploring different implemented methods for combining effects of multiple modest genetic variants at gene and pathway levels. We then propose a graph-based method that incorporates the signal from GWA study in the human protein-protein interaction (PPI) network to identify significant sub-networks or pathways associated with the trait of interest. This graph-based method applies centrality measures within linkage disequilibrium (LD) on the network to search for

significant sub-networks and applies a scoring summary statistic on the resulting sub-networks to identify the most enriched pathways associated with complex diseases. In addition, the proposed method also tests for possible signals of unusual differences in excess/deficiency of ancestry at gene and pathway levels in admixed populations.

This work starts with an introduction to genetic association studies and genome-wide association studies, which contains three parts. First, we start with basic concepts on population-based association studies, focusing on the case-control design of these studies. Linkage disequilibrium, a very important concept in population studies that measure dependencies between loci on the same chromosome, is then introduced. We also discuss two measures of genetic risks in case-control studies (odds ratio and relative risk), and their inference. One major pitfall of case-control association studies is hidden population substructures that may not have been considered. Basic definitions of two common population substructures, and methods for detecting and correcting population stratification and cryptic relatedness are introduced. The concept of a genome-wide association study is introduced, and their design, current limitations and future directions are broadly discussed.

Chapter 2 introduces pathway-based approaches for GWA study. This chapter is really important in the context of this study, as it introduces basic concepts on pathway-based analysis of GWAS datasets. It first covers resources needed to exploit the pathway-based approach for GWA studies such as protein-protein interaction (PPI) and pathway annotation databases, before exploring the motivation and different existing methods for pathway-based analysis of GWAS datasets.

Chapter 3 presents a new proposed method developed here, ancGWAS, for examining disease association signals from multiple interacting genes. It uses an algebraic graph-based approach, by integrating disease association signals from single SNPs, and the locus-specific ancestry information into the human PPI network to identify enriched sub-networks for the disease, and testing for any signal of excess or deficiency in ancestry. Different methods for mapping SNPs to genes, combining individual p-values and ancestral information at the gene level, constructing the LD-weighted PPI network, searching algorithms for network clustering, sub-network scoring and test for differences in ancestry at gene and sub-network levels, and finally pathway enrichment of identified modules are presented. Finally, a short note on the implementation of the new model is described in a detailed work-flow, and some concluding remarks are presented.

A thorough evaluation of the new proposed method is performed in Chapter 4, using 2 simulated case-control datasets, a pathway-based GWAS dataset for a heterogeneous population, as well as simulated interactive disease loci in a complex admixed population. We compare ancGWAS to an existing pathway-based method, dmGWAS, to assess the ability of ancGWAS to approximate the disease pathway. Finally, we apply ancGWAS to real data from the Cancer Genetic Markers of Susceptibility (CGEMS) for postmenopausal women of European ancestry

with invasive breast cancer with $1,145$ cases and $1,142$ controls genotyped at $528,169$ SNPs. The results are presented and discussed broadly including new biological insights that arose.

chapter 1

Introduction: Genetic association studies and Pathway-based approaches for GWAS analysis

The advances in genotyping technologies allowing for large-scale sequencing efforts have increased the power to uncover the genetic underpinnings of complex diseases, driving a multitude of investigators to embark on population-based genetic association studies. These studies differ from traditional epidemiological studies by their explicit considerations of genetic factors and family resemblance, instead of just focusing on the relationship between the environment and the occurrence of the disease, and its determinants in populations (Burton, P. R. *et al.*, 2005; Morton, N. E., 1982). Genetic variations occur in populations all the time, some with harmful effects contributing to an increase or decrease in disease risks in those populations. Some of these variations can reach substantial frequency in the population due to random drift or just by natural selection and they can exist as many different types depending on their effect on protein function, the associated susceptibility to a disease and their physical nature (Burton, P. R. *et al.*, 2005; Morton, N. E., 1982). In the latter case for instance, when a variation occurs at a single location within a gene and is present in at least 1% of a population, it is commonly referred to as a single-nucleotide polymorphism (SNP). Humans share common ancestry, and with the completion of the HapMap project along with the availability of genotyping technologies, enabling the genotyping of more than 1million genetic polymorphisms at once, as well as the genome-wide characterization of the levels and patterns of these human variations, genetic association studies of complex traits have become more comprehensive and effective.

Here, we discuss how genetic association studies are performed to determine whether these genetic variants are associated with a disease, focusing on case-control studies, in attempt to answer the second of two essential questions in genetic association studies including: (1) is there a genetic component to the disease, and (2) what genes are involved? Special attention will be

given to genome-wide association (GWA) studies, which involve whole and partial genome-wide scans in order to identify associations between SNPs and a trait, in order to identify genetic risks factors that can later be thoroughly analyzed using classical epidemiology methods. Population substructure is discussed, as well as different existing approaches to control for confounding due to population substructure for GWA study among population-based samples (Khoury, M. J. *et al.*, 1993; Lunetta, K. L., 2008). We also introduce PA approaches for analysis of GWA studies. We embark on discussing the current state of available biological knowledge on genes, which are generally exploited in these PA approaches. This biological knowledge includes the Gene Ontology (GO), and known human protein-protein interaction (PPI) data. We only focused on human PPI data, generally freely available in online databases. We also present some technical differences among current PA approaches, as well some common challenges observed in these methods.

1.1 Population-based association studies

In the context of genetic association studies, a phenotype, defined formally as a physical attribute or the manifestation of a trait, refers to a measure of disease progression. A phenotype can either be quantitative or binary, which refer respectively to continuous and binary variables. A quantitative trait refers to one that can often be measured and given a quantitative value whereas a binary trait is one that can take on two values, such as healthy or unhealthy. In this last case, the phenotype can easily be predicted knowing the genotype (Andrea, S. F., 2009). Clinical measures such as height and total cholesterol level are examples of quantitative traits, whereas a heart attack as an indicator for a cardiovascular outcome and the indicator of whether or not a patient has tuberculosis are examples of binary traits. Genetic association studies should only be undertaken if the trait of interest has established evidence for heritability, and when the question of whether a disease has a genetic component arises, the detection of higher occurrence rates in siblings or offspring is generally the first step. Other classical observational designs including studies on twins, adoptees and even migrants may suggest whether or not there is a genetic component in the etiology of a disease or trait of interest. Familial aggregation of a trait is a necessary but not sufficient condition to infer the importance of genetic susceptibility, because environmental and cultural influences can also aggregate in families, leading to family clustering and excess familial risk, as for instance similar environment may also contribute to familial aggregation (Tevfik, D. M., 2009).

Population-based association studies generally aim to relate genetic information to a clinical outcome or phenotype at a population level, in contrast with family-based studies. The latter involves data collected on multiple individuals within the same family unit. One statistical consideration that differs between family-based and population-based studies is that individuals in the same family are more likely to be similar to one another than individuals from differ-

ent families. That is, the trait under study is more likely to have a high correlation among individuals within the same family. In population-based studies instead, the assumption of independence between individuals is fundamental, and this non relatedness actually means that the relationship between individuals is unknown and assumed to be distant. Although there exist several situations where this assumption may be violated, such as in the case of repetition of measurements of a trait in study on the same individual; and in cases of within-cluster correlation, where for instance study samples may be collected across multiple hospitals, implying that patients from the same hospital are more likely to be similar than those across hospitals. In all of these cases, analytical methods for relatedness among the study individuals are essential for correctly estimating variance components (Andrea, S. F., 2009), described briefly in Section 1.2.2. Though there exist several study designs for genetic association studies, this review will focus specifically on the design and statistical analysis methods in the case-control study.

1.1.1 Case-control design

A typical design to test for population-based association is the case-control study, in which the frequency of SNP alleles in a series of well-defined groups of unrelated individuals are compared. Cases are those who have been diagnosed with the disease under study, and controls are a series of individuals who have been screened as negative for the presence of the disease in study or have been randomly selected from the population and must be at similar risk of developing the disease (Lewis, C. M., 2002). A clear definition of cases and controls is very crucial in a case-control study, and the same data must be collected the same way from both groups otherwise the study may suffer from researcher bias. Though sometimes less valued for being retrospective, case-control studies are cost-effective and they are actually the only ethical and efficient way to investigate association between an exposure and an outcome (Andrea, S. F., 2009). If carried out with care, with the definition and selection of controls, and reducing potential biases, case-control studies can generate effective results. A higher frequency of a SNP allele in cases compared with controls can be interpreted as the SNP allele increasing risk of disease, although several other interpretations can be considered (Section 1.1.3).

1.1.1.1 Statistical analysis methods of case-control studies

Let us consider an analysis on a single SNP with alleles A and B on a set of cases and controls. The data generated by this design can be summarized in a 2×3 or 2×2 contingency table consisting of six or four counts of the numbers of genotypes ($G_0 = AA$, $G_1 = AB$ and $G_2 = BB$) for each group. Assume that a total of r cases and s controls need to be tested, and n_0, n_1 and n_2 are the total number of AA, AB and BB observed genotypes, respectively. The genetic model (Lewis, C. M., 2002), which refers to a specific mode of inheritance, has to be determined in order to decide which is the right statistical test to use for association analysis. This

will also determine whether a genotype-based approach or an allele-based approach (additive or multiplicative models) must be undertaken (Zheng, G. *et al.*, 2012). The general genetic model (case in Table 1.1(a)) requires 2 degrees of freedom (*df*) for testing an association with a SNP, whereas the other models require only 1 *df*. The general genetic model retains the 3 distinct genotypes AA, AB and BB, with no assumptions about how the risk or mean for heterozygotes compares with the 2 homozygotes (Lunetta, K. L., 2008). Depending on the genetic model, a particular test statistic can be applied to this contingency table (see Table 1.1). There exist several test statistics to analyze association between the genotype and the disease under study, generally genotype-based and allele-based approaches. If the Hardy-Weinberg equilibrium[1] holds, both approaches are equivalent, otherwise allele-based tests can generate invalid results (Freidlin, B. *et al.*, 2002). Here, we will focus on the genotype-based approach using goodness-of-fit tests, including the Pearson's chi-squared test and the Cochran-Armitage Trend Tests (CATT), which are the commonly used test statistics for genetic association using a case-control design, rather than likelihood-based or regression methods and the allele-based test. When the Hardy-Weinberg holds, the trend test is based on the assumptions that one of the alleles is the risk allele, so that the risk of developing the disease is proportional to the number of risk alleles in the genotype. The trend test may not be powerful when the genetic model is misstated, as different trend tests are used for the four different genetic models (see Table 1.1(b-e)).

For example, with the hypotheses that allele B increased disease risk (dominant model), the case in (b) will be considered with a 2×2 table. On the other hand, if a recessive model is considered, two copies of allele B will be required to increase disease risk with a 2×2 table too. In a multiplicative model, whereby observed genotypes are broken down to compare the frequency of A and B alleles in cases and controls, the disease risk is increased by a factor r for genotype AB and of $2r$ for genotypes BB. This model shows a clear trend of an increased number of AB and BB genotypes, with disease risk for AB genotypes approximately half that for BB genotypes, whereas in an additive model, the disease risk is increased by a factor r for genotype AB and of r^2 for genotypes AA and BB. The latter can be tested using Armitage's test for trend (Zheng, G. *et al.*, 2012).

1.1.1.2 Pearson's Chi-Squared Test

The Pearson's test instead does not rely on any genetic information, thus it remains robust to any given genetic model. It is generally used for assessing deviation from the null hypothesis that cases and controls have the same distribution of genotype counts, and follow a chi-squared

[1]The Hardy-Weinberg principle predicts how gene frequencies will be inherited from generation to generation given a specific set of assumptions. It states that in a large randomly breeding population, allelic frequencies will remain the same from generation to generation assuming that there is no mutation, gene migration, selection or genetic drift (Lewis, C. M. *et al.*, 2012).

Table 1.1: Case-control study contingency tables for genetic models. Test (a) is the starting point for any analysis unless prior hypotheses are made. r_0, r_1, r_2, s_0, s_1 and s_2 are observed genotype counts in cases and controls.

(a) Full genotype table for general genetic model

	AA	AB	BB	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

(b) Dominant model: allele B increases disease risk

	AA	$AB + BB$
Cases	r_0	$r_1 + r_2$
Controls	s_0	$s_1 + s_2$

(c) Recessive model: two copies of allele B required to increase disease risk

	$AA + AB$	BB
Cases	$r_0 + r_1$	r_2
Controls	$s_0 + s_1$	s_2

(d) Multiplicative model: r-fold increased disease risk for AB and r^2 for BB. Tested only using allelic approach.

	A	B
Cases	$2r_0 + r_1$	$r_1 + 2r_2$
Controls	$2s_0 + s_1$	$s_1 + 2s_2$

(e) Additive model: $r-$fold increased disease risk for AB and $2r$ for BB. Analyzed only using genotypic approach.

	AA	AB	BB
Cases	r_0	r_1	r_2
Controls	s_0	s_1	s_2

distribution even when the Hardy-Weinberg proportion does not hold in the population (Lewis, C. M. *et al.*, 2012).

We assume the general genetic model described in Table 1.1(a), with (r_0, r_1, r_2) and (s_0, s_1, s_2), the counts for (AA, AB and BB) genotypes, and the total number is $r = r_0 + r_1 + r_2$ and $s = s_0 + s_1 + s_2$ respectively in cases and controls, with their expected values. Let the total number of genotype G_j be $n_i = r_i + s_i$, for i in $\{1, 2\}$, and $n = r + s = n_0 + n_1 + n_2$ be the total number of individuals. Under H_0, the null hypothesis that the disease status and genotypes are independent, the expected counts for cases r_i is given by $E_{r_i} = {}^{n_i r}/_{n}$ and for controls s_i by $E_{s_i} = {}^{n_i s}/_{n}$. The Pearson's test T is given by

$$T_{\chi^2_2} = \sum_{i=0}^{i=2} \frac{(r_i - E_{r_i})^2}{E_{r_i}} + \sum_{i=0}^{i=2} \frac{(s_i - E_{s_i})^2}{E_{s_i}} \sim \chi^2_2 \tag{1.1}$$

$T_{\chi^2_2}$ asymptotically follows a chi-squared distribution χ^2_2 with 2 *df*. H_0 is rejected at the level α if $T_{\chi^2_2} > \chi^2_2 (1 - \alpha)$, where $\chi^2_2 (1 - \alpha)$ is the $100 (1 - \alpha)$th percentile of χ^2_2.

Let us consider a single marker (*rs1219648*) in the *FGFR2* gene associated with risk of sporadic postmenopausal breast cancer from the CGEMS Breast Cancer study (Hunter, D. J. *et al.*, 2007). The contingency Table 1.2 represents genotype counts for different genotypes on this data, with one missing genotype in controls. This table shows that cases have higher frequency of both *GG* and *AG* genotypes compared with controls.

Table 1.2: *rs1219648* genotypes for a breast cancer case-control study with 1,142 controls and 1145 cases

		Genotypes			
Cohort	No. of individuals	*AA*	*AG*	*GG*	Frequency of allele G
Cases	1,145	352 (30%)	543 (47%)	250 (21%)	40%
Controls	1,142	433 (37%)	538 (47%)	170 (14%)	45%
Total	2,287	785	1,081	420	

We can determine whether the risk of developing the disease is proportional with the number of the alternate allele G by estimating the probabilities of having allele G given disease statuses by $\hat{\Pr}(G \mid \text{case}) = (r_1 + 2r_2)/(2r) = (543 + 2(250))/2287 = 0.456$ and $\hat{\Pr}(G \mid \text{control}) = (s_1 + 2s_2)/(2s) = (538 + 2(170))/2287 = 0.384$. $\hat{\Pr}(G \mid \text{case}) > \hat{\Pr}(G \mid \text{control})$, therefore the allele G is likely to be the disease risk allele.

Applying $T_{\chi_2^2}$ on Table 1.2 with the observed genotype counts in cases and in controls respectively $(r_0, r_1, r_2) = (352, 543, 250)$ and $(s_0, s_1, s_2) = (433, 538, 170)$, the expected genotype counts of AA, AG and GG for cases and controls are respectively 393.19, 541.45 and 210.37. Under the null hypothesis "H_0: There is no association between the marker and the disease", and the alternative hypothesis "H_1: There is association between the genetic marker and the disease", we can test for this hypothesis using the Pearson's chi-squared test T by

$$T_{\chi_2^2} = \frac{(352 - 393.19)^2}{393.19} + \frac{(543 - 541.45)^2}{541.45} + \frac{(250 - 210.37)^2}{210.37} + \frac{(433 - 393.19)^2}{393.19} + \frac{(538 - 541.45)^2}{541.45} + \frac{(170 - 210.37)^2}{210.37}$$

$T_{\chi_2^2} = 23.6123$ with p-value $= 7.459 \times 10^{-06}$ with 2 df. This suggests that the distribution of this data is due entirely to chance, with a 0.0007% chance of finding a discrepancy between observed and expected genotype distribution, disproving the null hypothesis of no association between the marker M and the phenotype. Thus, we can conclude the existence of a moderate association between the marker *rs1219648* and breast cancer through this simple test statistic using the Pearson's Chi-Squared Test. More specifically, genetic associations can be measured using the odds ratio (see Subsection 1.1.2), whereby the null hypothesis of no association between the marker and the disease can be defined as H_0: $\theta = 0$ and the alternative hypothesis as H_1: $\theta \neq 0$. If the direction of the association is specified and the risk allele is known, a

one-side alternative hypothesis can used and defined as H_1: $\theta > 0$ or H_1: $\theta < 0$.

1.1.2 Measure of genetic risks in case-control studies

After assessing for departure of the distribution of SNP allele frequencies in cases and in controls, estimating the disease risk conferred by the SNP allele is a decisive routine in genetic association studies. These measures estimate the effect size of the SNP allele with respect to the disease, instead of a significance value (John, F. Y. B., 2010). These are also useful if the time lag between the exposure and the disease is very long or if the disease of interest is rare.

Table 1.3: 2×2 table of levels of risk factor for n individuals with case and control status.

	$E+$	$E-$	
Case	a	b	$a+b$
Control	c	d	$c+d$
	$a+c$	$b+d$	n

In a cohort study design, where a group of individuals exposed to a disease agent and a well selected group of individuals without the disease and not exposed to the disease agent (Table 1.3) are observed and compared until an event of interest occurs or for a specified period of time, the association between the exposure and the disease can be expressed as the relative risk (RR), which is regarded as the proportion of being exposed versus not being exposed, defined as

$$RR = \frac{a\,(a+b)}{c\,(c+d)}.$$

In a case-control study design on the other hand, the Odds Ratio (OR) for a disease is a commonly used measure of association in genetic association studies, which compares the odds that an outcome will occur in the presence a specific exposure to the odds of the outcome occurring in the absence of that exposure, or simply the ratio of allele porters to the non-porters in cases compared with the same ratio in controls (Lewis, C. M., 2002).

Technically, even though easy to interpret, as the apparent RR is dependent on the number of controls chosen and the disease prevalence, RR should not be used to express results in case-control studies. However, the OR, though less intuitive, can be a reasonable approximation of the RR when the disease is relatively rare or when less than 1% of people exposed to an agent develop disease. When the disease rate rises above 1% it produces progressively larger errors, in this case, when the study design allows for calculation of a RR, OR can be interpreted as if it were RR, comparing risks rather than just a comparison of odds.

For a retrospective case-control study, we define the odds of being $E+$ (exposed) versus being

$E-$ (not exposed) as

$$\frac{\Pr\left(E+ \mid d\right)}{\Pr\left(E- \mid d\right)}, \tag{1.2}$$

where $d = 1$ is for a case and $d = 0$ is for a control. Thus, the OR with respect to two levels of R, e.g. with respect to case and control groups is defined as

$$\mathrm{OR}_{R=E+:R=E-} = \frac{\Pr\left(E+ \mid d=1\right)}{\Pr\left(E- \mid d=1\right)} \Big/ \frac{\Pr\left(E+ \mid d=0\right)}{\Pr\left(E- \mid d=0\right)} \tag{1.3}$$

Considering the 2×2 table given in Table 1.3, the estimate of the OR is given by

$$\widehat{\mathrm{OR}} = \frac{ad}{bc}, \text{ or } \log \widehat{\mathrm{OR}} = \log\left(\frac{ad}{bc}\right), \tag{1.4}$$

where the log transformation is used to promote normality as the odds ratios do not follow a normal distribution, which can be used to produce the confidence interval for the OR. The standard error of the $\log \widehat{\mathrm{OR}}$ can estimated as

$$\mathrm{SD}_{\log \widehat{\mathrm{OR}}} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}, \tag{1.5}$$

referred tp as the Woolf's estimate of the standard error (Lewis, C. M. *et al.*, 2012), and the confidence interval (CI) for the $\log \hat{\mathrm{OR}}$ can be obtained from

$$e^{\log \widehat{\mathrm{OR}} \pm z \mathrm{SD}_{\log \hat{\mathrm{OR}}}}, \tag{1.6}$$

where $e \approx 2.71828$ is the base of the natural logarithms, and z is a Standard Normal deviate corresponding to the level of confidence (1.645 for 90% confidence, 1.96 for 95% confidence, and 2.576 for 99% confidence).

In the case of a diallelic marker with alleles A and B as described in the 2×3 contingency table in Table 1.1, two ORs can be derived to measure the association. Let us denote OR_1 the OR between AA and AB and OR_2 the OR between AA and BB, using the genotype with wild-type AA homozygotes as a reference in both cases. These OR_i and their asymptotic standard deviation are given by

$$\log \widehat{\mathrm{OR}}_i = \log\left(\frac{r_i s_0}{r_0 s_i}\right), \ \mathrm{SD}_{\log \widehat{\mathrm{OR}}_i} = \sqrt{\frac{1}{r_0} + \frac{1}{r_i} + \frac{1}{s_0} + \frac{1}{s_i}}, \text{ for } i \text{ in } \{1, 2\}. \tag{1.7}$$

If we assume an unbiased and well conducted study: an OR of 1 implies no association between the exposure and the outcome; an OR of more than 1, and if its 95% CI does not include 1, is said to be statistically significantly different, that is, a positive association between the exposure and the outcome at the 5% significance level. In other words, the odds of exposure are greater in cases than in controls, which suggests in the context of genetic association studies

that the marker or variant increases the risk of the disease. An OR of less than 1, and if its 95% CI does not include 1, implies a negative association between the exposure and the outcome at the 5% significance level. In other words, the odds of exposure are greater in controls compared to cases, suggesting a protective role of the exposure with respect to the outcome. If the 95% CI contains 1, even though the resulting OR is greater or less than 1, the association between the exposure and the outcome is not proven at the 5% significance level (Szumilas, M., 2010; Bigby, M., 2000).

The genotype data for the marker *rs1219648* in Table 1.2 was found to have significant association with the disease under study from Pearson's chi-square test with p-value$= 7.459 \times 10^{-06}$. Computing the odds ratio for this association as described in Equation 1.7, we have $\text{OR}_1 = 1.24$ with 95% CI $= (1.03, 1.50)$ and $\text{OR}_2 = 1.18$ with 95% CI $= (0.93, 1.50)$. From these results, we observe evidence for association between the marker *rs1219648* and breast cancer in homozygote cases, however it does not reach statistical significance in heterozygote cases, as the 95% CI of the OR for this group of individuals spans 1.

1.1.3 Linkage disequilibrium and indirect association

Case-control study design generally requires large numbers of cases with the disease and controls with thousands to millions of SNPs genotyped using currently available chip-based microarray technologies, taking into account both sample size and SNP coverage to have sufficient power. SNPs may be selected across the genome or may be selected specifically for their coverage using existing data, which can be used as tagging SNPs. Thus some considerations have to be taken into account to ensure that the majority of required loci and individual sites that must be examined are captured. First, the specific genetic variation differences in populations must be determined to enable the proper study design to be undertaken for different populations. Secondly, the position and density of commonly occurring variations in the population need to be known in order to avoid or reduce redundancy, taking into consideration the existing correlations among SNPs to be captured (Bush, W. S. *et al.*, 2012). As a result of the association between these SNPs or *linkage disequilibrium*, they can be considered as markers, helping to locate genes that are associated with the disease.

1.1.3.1 Linkage disequilibrium

Linkage disequilibrium (LD) refers to the property of SNPs in the same region of the genome that describes the degree to which an allele of a SNP M_1 is correlated or inherited from an allele of another SNP M_2 within a population. This dependence between loci on the same chromosome pair and close enough to each other arises because any novel locus occurs on a background of fixed alleles at other loci (Lewis, C. M. *et al.*, 2012). Suppose the SNP M_1 is a

functional locus for a disease with unobserved alleles, and the SNP M_2 is a non functional locus with observed alleles, M_2 is said to be a marker. Let us assume A and B the normal alleles, and a and b the disease alleles for M_1 and M_2, respectively, with allele frequencies given by $\Pr(a) = p$, $\Pr(b) = q$, $\Pr(A) = 1-p$ and $\Pr(B) = 1-q$. We define the linkage disequilibrium coefficient for each haplotype, under independence by

$$D = \Pr(AB) - \Pr(A)\Pr(B), \tag{1.8}$$

as the deviation between the observed and expected numbers of disease alleles and normal alleles or its equivalent expressions $\Pr(ab) - \Pr(a)\Pr(b)$, $\Pr(AB) - \Pr(A)\Pr(b)$, $\Pr(BA) - \Pr(a)\Pr(B)$, $\Pr(AA)\Pr(BB) - \Pr(Ab)\Pr(aB)$. The joint distribution of the alleles at the two loci M_1 and M_2 are given in Table 1.4.

Table 1.4: 2×2 joint distribution of marker and functional locus in the same region of the genome for a sample from a given population.

	B	b	
A	$(1-p)(1-q)+D$	$(1-p)q - D$	$1-p$
a	$p(1-q) - D$	$pq - D$	p
	$1-q$	q	1

A D value of 0 indicates the two loci are independent and are said to be in complete linkage equilibrium, implying frequent recombination and statistical independence under principles of Hardy-Weinberg equilibrium (Bush, W. S. *et al.*, 2012), whereas when $D = 1$, the two loci are in complete LD, implying no recombination between the two loci. When the marker SNP is also a functional SNP and their allele frequencies are the same, we have the following cases:

- Single-locus model: $A = B$, $a = b$ implying $p = q$, $\Pr(AB) = \Pr(A) = 1-p$ and $D = p(1-p)$;
- Two-locus model: $A = b$, $a = B$ with $p = 1-q$, $\Pr(AB) = 0$ and $D = -p(1-p)$. This model also implies when $D \neq 0$, the two loci are also said to be in gametic phase disequilibrium.

Mutations arise in a population initially closely associated with existing polymorphisms, that is in high LD with neighboring SNPs and they are transmitted as a haplotype (see Figure 1.1). If a mutation is causative with strong disease risk, then SNPs in the region can be used as markers of its presence. These SNPs usually maintain and appear to be in strong statistical association with the disease of interest. Two commonly used measures of linkage disequilibrium are the standarized LD coefficient D' and the squared Pearson coefficient of correlation r^2, standardized in both case to remove the arbitrary sign introduced by allele

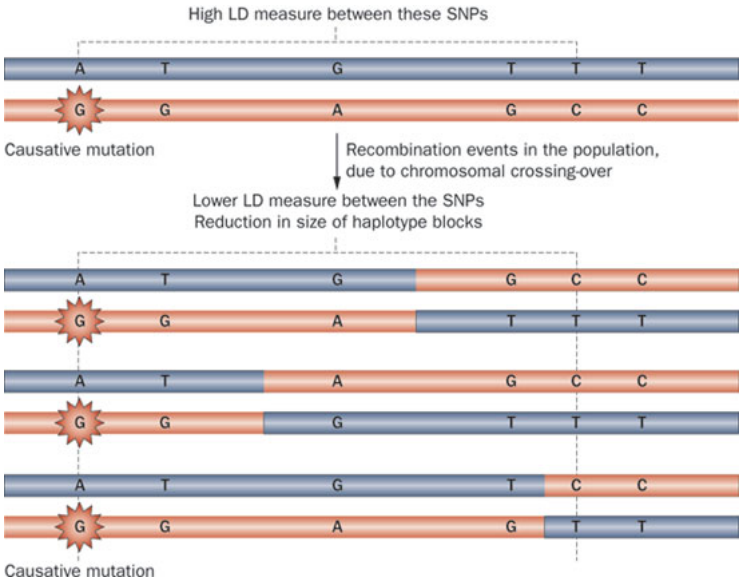

Figure 1.1: Two SNPs are initially in strong LD in a population, but as mutations spread within the population through recombination events over generations, LD between these two SNPs breaks down. During this process, the sizes of initial haplotype blocks are shrunk, giving rise to six different haplotypes in the population. Over time, a pair of SNPs move from linkage disequilibrium to linkage equilibrium. Reproduced from (Rosset, S. *et al.*, 2011).

frequency of two loci (Devlin, B. *et al.*, 1995), defined by

$$D^{'} = \begin{cases} D/\min\{(1-q)\,p, (1-p)\,q\} & if D > 0. \\ D/\min\{(1-p)\,(1-q)\,, pq\} & if D < 0. \end{cases} \tag{1.9}$$

$$r^2 = \frac{D^2}{pq\,(1-p)\,(1-q)}. \tag{1.10}$$

Though $D^{'}$ relies on recombination events, there exist some dependencies between these two measures, in the sense that r^2 is sensitive to the allele frequencies of the two SNPs and can only be high in regions of high $D^{'}$. In these cases, $|D^{'}| = 1$ refers to complete LD and $|D^{'}| = 1$ together with $A = B$ and $a = b$ with $p = q$ $(D > 0)$, $A = b$ and $a = B$ with $p = 1 - q$ $(D < 0)$ refers to perfect LD (Zheng, G. *et al.*, 2012).

LD varies across human sub-populations and genome regions and between pairs of markers in close proximity. The degree and pattern of LD is also influenced by many factors on several levels of comparison. Factors such as genetic drift, population size, admixture and inbreeding are specific to a population, whereas factors such as natural selection, gene conversion and recombination rate are specific to the genomic region (Shifman, S., 2003). Thus, African populations have shorter regions of LD as they are the most ancestral and have accumulated more recombination events (Figure 1.1a). European-descent and Asian-descent populations, which were created by sampling of chromosomes from the African populations, resulting in change in the number of founding chromosomes, and size, as well as the generational age of the population, have, on average, longer regions of LD in contrast to African populations (Figure 1.1b).

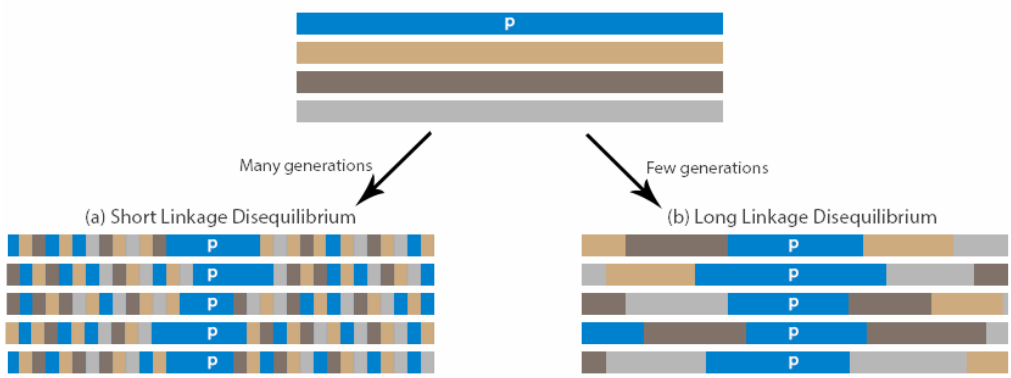

Figure 1.2: The disease polymorphism (p) is shown in white and the adjacent chromosomal markers are shown in their chromosome colour accordingly. Due to recombination, they have been shuffled. **(a)** Mutations that occurred recently have had fewer recombinational events, therefore they are associated with long LD intervals. **(b)** Mutations that occurred a long time ago have had many more recombinational events, therefore they are associated with short LD intervals. Adapted from (Ostrer, H., 2001).

To further illustrate what the impact of LD differences between populations can be in the design of case-control genetic association studies, we consider the *WT1* gene that was recently identified to be associated with a protective role in the pathogenesis of tuberculosis (TB) (Thye, T. *et al.*, 2012; Chimusa, E. R. *et al.*, 2013), located in the p13 region on chromosome 11. Figure 1.3 illustrates how LD structures differ in different populations. Thus, as the patterns of LD vary in populations, the ability to detect association within the region harboring the disease allele will depend on the strength of LD between unobserved and observed markers accordingly, consequently influencing the specific trait-associated markers.

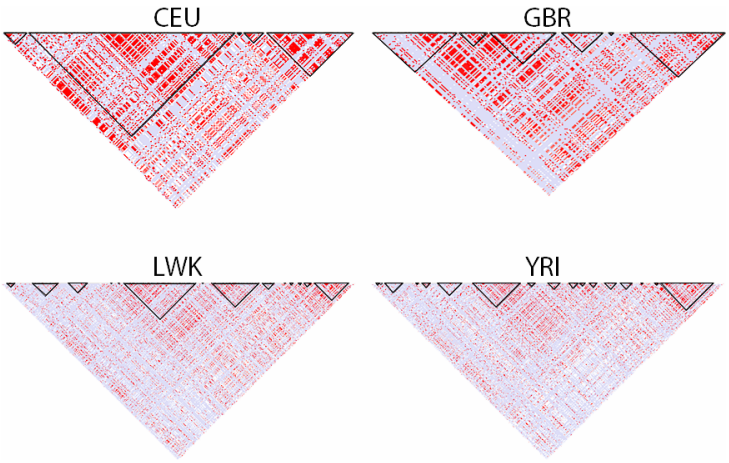

Figure 1.3: LD patterns of the *WT1* gene in different populations from the HapMap 3 project. The Utah Residents (CEPH) with Northern and Western European ancestry (CEU), and the British in England and Scotland (GBR) represent two European populations with generally large blocks of LD, whereas the Luhya in Webuye in Kenya (LWK) and the Yoruba in Ibadan in Nigeria (YRI) representing African populations with small blocks of LD.

1.1.3.2 Direct and indirect association

Because the functional locus that has a causal effect for a disease is unknown, a marker is typed and subsequently tested for association with the disease. Figure 1.4 describes the three outcomes that are possible from the subsequent genetic association study. In the first outcome, the polymorphism is a putative causal variant. This case is referred to as a *direct association*, in which the variant influencing a biological system, which in the end leads to the disease, is genotyped and found to be statistically associated with the phenotype in the study (Bush, W. S. *et al.*, 2012). Exomic polymorphisms, which cause change in an amino acid, are generally considered as candidate variants for association studies, but it is also possible that mutations that occur in non-coding regions influence the pathogenesis of complex diseases. This type of study is easier to carry out and the most powerful, the difficulty resides in the identification of candidate variants, as we actually do not know which variants might have effects on the disease, or which part of the genome harbors functional polymorphisms (Cordell, H. J. *et al.*, 2005).

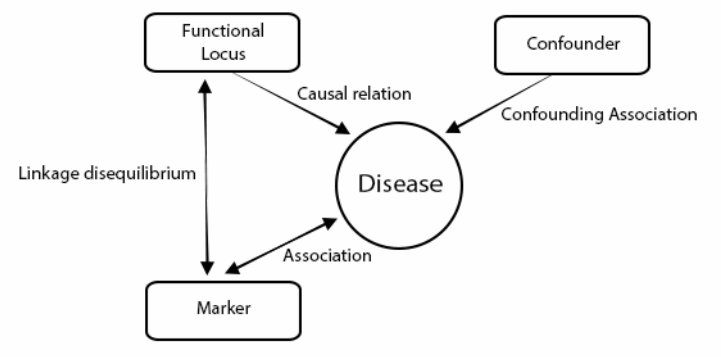

Figure 1.4: Three forms of association between a marker and a disease and the relationship among the marker, the functional locus, the confounder and the disease.

In the second situation, the functional marker is not genotyped, instead a polymorphism that is statistically associated with the disease is typed and is in high LD with the influential marker (Figure 1.5). This polymorphism acts as a proxy to the causal polymorphism and this case is referred to as an *indirect association*. Indirect associations are less powerful than direct associations and they are even more difficult to analyse (Cordell, H. J. *et al.*, 2005). As in most cases, indirect associations have weaker signals than the causal associations they reflect, so as many markers as possible surrounding the causal marker need to be genotyped to have a higher chance of detecting those indirect associations (Figure 1.5).
Considering these first two possibilities, a statistically significant SNP association from a case-control study should not be considered as the causal polymorphism unless additional studies are carried out to determine whether the identified SNP has an indirect or direct effect on the disease, or just has a surrogate role for a causal variant. One then needs to map the precise location of the casual variant. To this end, most of the time investigators concentrate on candidate genes identified either from animal model studies or on the basis of their known biological

function (Bush, W. S. *et al.*, 2012; Cordell, H. J. *et al.*, 2005).

Statistical association between a marker and the disease can also be an outcome of neither a direct or indirect association, but rather a result of false findings (positive confounding) or a result of a negative confounding factor obscuring the true causal association signals (Figure 1.4). In this third situation, the marker association might be due to the effects of chance (random error) or to confounding by stratification and substructure within the population, leading us to conclude the existence of a valid statistical association when one does not exist (false positive or type I error) or alternatively the absence of an association when one is truly present (false negative or type II error). Section 1.2 discusses different confounding factors that might disrupt case-control design studies including population stratification and cryptic relatedness, and some existing methods for detecting and circumventing this difficulty.

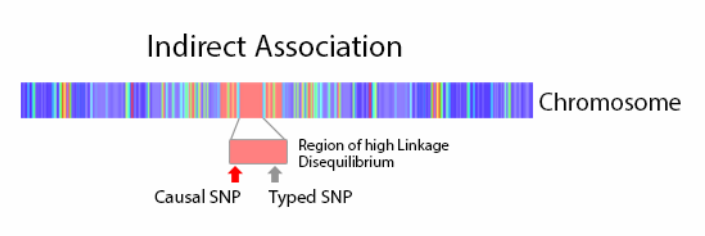

Figure 1.5: Typed SNPs can lie in a region of high linkage disequilibrium with a disease risk SNP. Markers should be selected specifically to capture the variations at nearby sites and serve as surrogates for functional SNPs through indirect associations.

Because of these different possible interpretations of an association test outcome, a concise knowledge of the variability of linkage disequilibrium is essential to appropriately developing and interpreting association studies. In addition, as the variance of linkage disequilibrium is also sensitive to population structure, each region included in a genotype-phenotype association study should be assessed accordingly (Goldstein, D. B. *et al.*, 2001).

1.2 Population structure and impact of population stratification

All mutations result in new alleles, which can persist or be lost in a population and their frequency can vary over the years due to chance or selection, or due to admixture between populations or be influenced by population size and bottlenecks. In the latter case, the population genetic variation can be reduced by a lot, resulting in the inability to adapt to new selection pressures, such as climatic change because the alleles that selection would act on may have already drifted out of the population (de Wit, E. *et al.*, 2010; Li, H. *et al.*, 2011; Caldwell, R. *et al.*, 2014). An example of this scenario is the South African Coloured (SAC) population, which is a result of mainly a few Dutch settlers, and has an unusually high frequency of certain alleles such as those involved in Huntington's disease or tuberculosis, just because those original colonists happened to carry those alleles with unusual high frequency (de Wit, E. *et al.*, 2010; Caldwell, R. *et al.*, 2014). If we consider a 20 year-generation population, with only

two or so average offspring per generation, a mutation spreads slowly for a long-time period, and this process is influenced by stochastic factors such as birth, death, and migration. A variation that occurs commonly in a population is likely to be relatively old even though it is a result of natural selection. Thus, it has disseminated over time via demic exchange over larger geographic areas, proportionally with frequency, age, and geographic distribution. Because of the recent history among species, common variation is typically global and those variants were probably present at the time of human expansion out of Africa approximately $100,000$ years ago.

The accumulation of these processes of mutation in populations for a long time period have shaped human variation, and because of Gregor Mendel's laws of inheritance, chromosomally linked markers and functional variants are co-transmitted to offspring generations, reflecting the history of shared genetic variation or "ancestry". As a result, markers are shared among samples of related individuals with statistically specifiable patterns of disease (Kittles, R. A. *et al.*, 2003). Investigators typically use this statistical information in trying to infer regions of the genome that harbor these shared disease markers, which can quickly lead to the gene. Investigators typically use this fact by first searching among closely related individuals. However, as related individuals share much of their entire genome, it can be advantageous to look among more distantly related groups of affected individuals who, because of recombination since their common ancestors, disproportionately share variation at the chromosome region responsible for the common disease (Kittles, R. A. *et al.*, 2003; Dries, D. L., 2009).

1.2.1 Population structure

Because all genes in a geographic region share the same general population history, there exists a greater correlation among variants found in one region compared to another geographic region. However, because of random drift or even selection, some non-linked genes might be subject to more or less independent changes in allele frequency resulting in genetic variant differences across a subpopulation (Zheng, G. *et al.*, 2012; Kittles, R. A. *et al.*, 2003).

It has been suggested that seven major clusters exist in the global human populations, determining the genetic structure of these populations, with the largest component lying along geographic lines. These clusters include the Europeans/West Asians (Whites), sub-Saharan Africans, North Africa/Middle Est, East Asians, Pacific Islanders, and Native Americans, where some subpopulations have mixed ancestries from other major geographic regions as illustrated in Figure 1.6A.

The SAC population of South Africa in Figure 1.6B is an example of a population with a mixed ancestry, which is generally interpreted as either a recent population admixture or shared ancestry before the divergence of contributing populations. It has been reported in accordance with historical data that the SAC population is rather a result of a recent admixture, whose

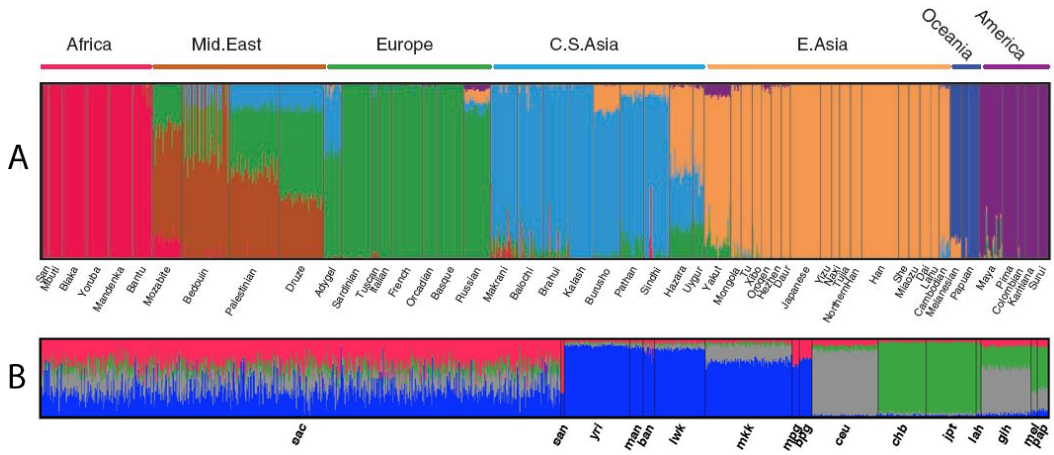

Figure 1.6: Proportion of an individual's ancestry. Each individual is represented by a vertical line. (A) Worldwide human population structure based on 650,000 SNPs from 938 unrelated individuals from 51 different populations using the frappe program. From (Li, J. Z. *et al.*, 2008). (B) South African Coloured (SAC) population based on 959 individuals with 75,000 autosomal SNPs genotyped from the Western Cape area of South Africa with individual ancestry generated using the STRUCTURE program. From (de Wit, E. *et al.*, 2010).

genomes consist of predominantly Khoisan $(32-43\%)$, Bantu-speaking Africans $(20-36\%)$, European $(21-28\%)$ and a smaller Asian contribution $(9-11\%)$, depending on the model used (de Wit, E. *et al.*, 2010) as in Figure 1.6B.
Therefore, the structure of a population can result in inbreeding because individuals in different subpopulations or ethnic groups can share common ancestries even in the occurrence of random mating. These substructures in population can affect case-control design for genetic association studies, and two types of substructure have been substantially considered in the literature, including population stratification and cryptic relatedness.

1.2.1.1 Population Stratification

Population stratification (PS) arises in case-control designs when two subpopulation groups in a population are poorly matched for genetic ancestry, resulting in a change in allele (genotype) frequency of the marker across the subpopulations (Lewis, C. M. *et al.*, 2012). Several genetic association studies have been cited as examples illustrating how population stratification can affect case-control studies (Thomas, D. C. *et al.*, 2002). Two circumstances have been identified as leading to this phenomena, therefore when testing for disease-marker association, their existence must be tested to confirm the influence of hidden PS on the study design. The two conditions which have to be satisfied in PS are whether; (1) the allele frequency of the marker varies across the subpopulations, and (2) the disease rate varies across the subpopulations.

To control the allele frequency of the marker for PS, one can just avoid large allele frequency differences between subpopulations or ethnic groups, which can be done using Wright's F-

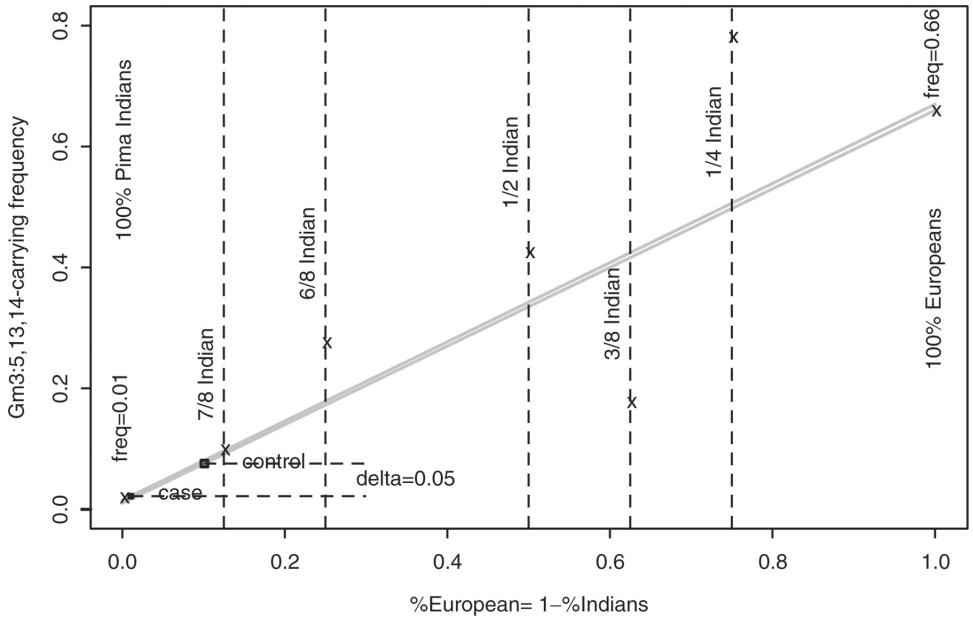

Figure 1.7: Illustration of PS effect in a candidate gene study of the potential association between $Gm^{3:5,13,14}$ and type 2 diabetes (T2D) using American Indian samples. An artifact of the haplotype derived from allotypic markers $Gm^{3:5,13,14}$-carrying frequency difference of 0.05 is produced by this unequal proportion of American Indian heritage in case and control groups. From (Li, W., 2008; Knowler, W. C. *et al.*, 1988).

statistic for the purpose of measuring the effect of admixing subpopulation (Wills, C., 2007). Let the allele frequency of an allele in n subpopulations be p_1 and p_2 for $n = 1, 2$, the allele frequency of another allele $q_1 = 1 - p_1$, $q_2 = 1 - p_2$, the frequency of heterozygote in the two subpopulations will be $2p_1q_1$, p_2q_2, and the heterozygote frequency in the combined population will then be

$$H_{\text{whole}} = 2\overline{pq}, \tag{1.11}$$

where $\overline{p} = wp_1 + (1 - w)\, p_2$, $\overline{q} = wq_1 + (1 - w)\, q_2$ are the averaged allele frequency of the two alleles, and w and $1 - w$ the mixing proportion of the two population, and the heterozygosity in each subpopulation is given by

$$H_{\text{sub}} = 2wp_1q_1 + 2p_2q_2\,(1 - w)\,. \tag{1.12}$$

We can observe that the frequency of heterozygotes always increases when the subpopulations are combined and can be written as $H_{\text{whole}} \geq H_{\text{sub}}$. Therefore, the inflation of the heterozygote frequency in the combined population will vary proportionally with the allele frequency difference $|p_1 - p_2|$. The percentage of increase of the heterozygote frequency by combining

subpopulations is the Wright's F=statistic given by

$$
\begin{aligned}
F \equiv \frac{H_{\text{whole}} - H_{\text{sub}}}{H_{\text{whole}}} &= \frac{2w\,(1-w)\,(p_1 - p_1)^2}{H_{\text{whole}}} \\
&\approx \frac{2w\,(1-w)\,(p_1 - p_1)^2}{H_{\text{sub}}} \\
&= \frac{(p_1 - p_1)^2}{(p_1 q_1)\,/\,(1-w) + (p_2 q_2)\,/\,(w)}
\end{aligned}
$$

While the F value can be marker-specific or dataset-specific, an $F = 10^{-1}$ appears to describe normal variation between major ethnic groups and an $F = 10^{-4} \sim 10^{3}$ appears to describe variation between regions of an isolated population (Li, W., 2008).

Figure 1.7 is a typical illustration in a genetic association study of type 2 diabetes (T2D), where a 10% difference in the composition of the studied population ethnic groups has lead to Type I error signal. In this study, carried out in 1988 by Knowler (Li, W., 2008), a haplotype $\text{Gm}^{3:5,13,14}$ is examined in American Indians with various degrees of Indian heritage. Assuming that the $\text{Gm}^{3:5,13,14}$-carrying probability is practically the same between the case and the control group, indicated by the two straight lines, the $\text{Gm}^{3:5,13,14}$-carrying frequency as a function of Indian heritage is shown in crosses (0 for European Caucasians, 7/8 for one non-Indian grand-grandparent, 1 for 100% Indian heritage and so on), we can observe a dramatic $\text{Gm}^{3:5,13,14}$-carrying frequency change with the population proportion with an increase from 0.01 for 100% Indians to 0.66 for 100% Europeans. The solid square indicates the case group with 99.9% Indian heritage, and the open square indicates the control group with 99% Indian heritage.

The occurrence of PS is also subject to population heterogeneity in disease rates. The allele frequency differences must also be correlated with the subpopulation differences in their baseline disease rates. If there is no variation in the rates not attributable to specific genes under study in the case of a candidate study for instance, then confounding cannot occur. Epidemiological studies on infectious diseases have stated that there exists a broad range of susceptibilities to major infectious diseases, depending on population history for the disease, and whether the populations have been exposed only relatively recently or the populations have been in equilibrium with respect to the disease for a long time (Thomas, D. C. *et al.*, 2002).

1.2.1.2 Cryptic Relatedness

While PS typically considers remote common ancestry of large groups of individuals, another source of confounding in a case-control design that might actually be a more serious source of error, Cryptic Relatedness (CR), refers to recent common ancestry among smaller groups

of individuals (Astle, W. *et al.*, 2009). In other words, CR states that some of the individuals in the case-control data might be close relatives of each other, resulting in their genotype not being drawn independently from the population frequencies. This could ultimately result in greater variance in the allele frequency estimates in cases and controls than expected even in the absence of allele frequency estimate biases, and generate inflated type-1 error when testing for genetic association using a test statistic this ignores that excess of relatedness among samples (Voight, B. F. *et al.*, 2005). Voight and Pritchard in (Voight, B. F. *et al.*, 2005) argued that CR may have negligible effects in well-designed studies of outbred populations, but may cause serious concerns in founder populations that have grown rapidly and recently from a small size. To demonstrate this, they analyzed a sample of six phenotypes from the Hutterite population. They showed that CR decreases an association signal of 10^{-3} by a factor of approximately 4, and that the smaller the signal the greater the relative effect. There might also exist a case where both PS and CR can have an effect on the case-control sample, where, in addition to relatedness among individuals, the allele frequencies and disease prevalences also change across subpopulations.

1.2.2 Correcting for population stratification and cryptic relatedness

In an association study, the sample population can be composed of individuals from different subpopulations or ethnic groups, whose allele frequencies as well as disease history, such as the levels of exposure, differ, as discussed in sections 1.2.1.1 and 1.2.1.2. This heterogeneity with respect to environmental factors among subpopulations or the genetic background may cause spurious correlations, leading to false-positive or false-negative results depending on whether these differences are in the same or opposite directions (Thomas, D. C. *et al.*, 2002).
Consider a genetic association study, testing for association between a genotype G at a particular locus and a phenotype Y, where individuals are collected from a continuous geographical space Z, directly correlated with PS. Let us assume a qualitative trait with disease status 0 and 1 for control and case, respectively. If we assume a conditional independence between the phenotype Y and genotype G of an individual given that the individual is sampled at a position $z \in Z$, there is no direct or indirect association between Y and G for a given z. In this case, the null hypothesis H_0 is equivalent to

$$\Pr(G|z, Y = 1) = \Pr(G|z, Y = 0) = \Pr(G|z), \tag{1.13}$$

where $\Pr(G|z, Y = 1)$ and $\Pr(G|z, Y = 0)$ are the genotype frequencies at z for a case and a control, respectively, implying no association conditional at the fixed region. If the specific regions where cases and controls are sampled from are unknown in an association study, the null hypothesis H_0 can be expressed as

$$\Pr(G|Y = 1) = \Pr(G|Y = 0), \tag{1.14}$$

If H_0 does not hold, based on (1.14) when (1.13) holds, the association is referred to as a *spurious association* Note that if cases and controls are collected from the same geographical region ($z \in Z$ is a constant) or from the same genetic background, then (1.13) implies (1.14) and a spurious association will not occur (Zheng, G. *et al.*, 2012).
PS and CR can act as confounding factors, which can be eliminated through better study design or detected and controlled in the analysis. When not eliminated through study design, statistical tests have to be adjusted for confounding factors that are known to have an impact on the trait such as age, sex, study site, and other clinical covariates. Covariate adjustment can extensively reduce spurious associations due to population structure and other sampling artifacts, and various methods can be used to protect association studies from confounding. Here, we discuss some major existing methods for detecting and adjusting for PS and CR in genetic association studies.

1.2.2.1 Kinship coefficients based on marker data

The relatedness between two individuals can also be defined as the probabilities that each subset of their alleles at any given locus is identical by descent (IBD), in other words, that they have been inherited from a common ancestral allele without an intermediate mutation (Astle, W. *et al.*, 2009). The correlation coefficient K_{ij} for variables indicating whether alleles drawn from each of i and j individuals are some given allelic type, let's say A, denotes the *Kinship coefficient*. Let us denote p the frequency of allele A, x_i and x_j the A allele counts of i and j respectively, with x_i and x_j in $(0, 1, 2)$, then K_{ij} can be estimated from genome-wide covariances of allele counts by

$$\mathrm{Cov}\left(x_i, x_j\right) = 4p\left(1-p\right)K_{ij}, \tag{1.15}$$

where p represents the population fraction of A alleles. Thus, the estimator of the kinship matrix K is given by

$$\hat{K} = \frac{1}{L}\sum_{l=1}^{L}\frac{\left(x_l - 2p_l\mathbf{1}\right)\left(x_l - 2p_l\mathbf{1}\right)^T}{4p_l\left(1-p_l\right)}, \tag{1.16}$$

where x is a column vector over individuals, p_l is the frequency of allele A at locus l and L is the number of loci. Entries in $\hat{K}$ can also be interpreted in terms of excess allele sharing beyond that expected for unrelated individuals, given the allele fractions. Although the correlations resulting from shared ancestry are in principle positive, because of bias resulting from estimation of the p_l, off-diagonal entries of (1.16) can be negative. However, for the purpose of phenotypic correlations, genotypic correlations seem to be intuitively appropriate. In this case, $\hat{K}_{ij}$ can be interpreted as excess allele-sharing, where negative values represent the sharing of fewer alleles than expected given the allele frequencies (Astle, W. *et al.*, 2009).

On the other hand, if we consider the probability that alleles chosen at random from each of two individuals match at a genotyped dialelic locus, that is, *identical by state* (IBS), the genome-wide average IBS probability can be written as

$$\frac{1}{2L}\sum_{l=1}^{L}\left(x_l-1\right)\left(x_l-1\right)^T+\frac{1}{2}. \tag{1.17}$$

IBS arises as a result of IBD, usually when the mutation rate is low. The excess allele-sharing or genotypic correlation estimator of kinship coefficients (1.16) is typically more precise than (1.17) because it includes weighting by allele frequency. In addition, because the rare allele is likely to have arisen from a more recent mutation event, sharing a rare allele suggests closer kinship than sharing a common allele (Astle, W. *et al.*, 2009; Slatkin, M., 2002).

The kinship estimate practically tells us how similar the genomes of people involved are, and is used in many methods in genetic association studies for protecting from confounding factors. Those methods are generally formulated within standard regression models in which, for the phenotype of the ith individual, the expected value y_i is expressed as a function of its genotype x_i at the SNP of interest as well as optional covariates such as gender or age as

$$g\left(\mathbb{E}\left[y_i\right]\right)=\alpha+x_i\beta, \tag{1.18}$$

where g is a link function and β is a scalar or column vector of genetic effect parameters at the SNP, and α is a normally distributed noise term that accounts for unexplained variation in y. In a case-control design, g is the logit function and β a matrix of log odds ratios. Here, case-control status is treated as the outcome, this refers to a prospective model. However, if considering a retrospective case-control model, where some ascertainment effects can not be correctly modeled prospectively, the standard linear model (1.18) can be expressed as

$$g\left(\mathbb{E}\left[x_i\right]\right)=\alpha+y_i\beta, \tag{1.19}$$

where g here is typically the identity function $g\left(y_i\right)=y_i$.

1.2.2.2 Using better measures of populations through study design

As mentioned earlier, confounding factors can be controlled and eliminated during study design, by matching, stratification, or multivariate adjustment models. In the case of population stratification for instance, the need for more detailed information on ethnicity than just broad conventional categories such as African, Caucasians, Asians and others, can be substantial. Thomas and Witte in (Thomas, D. C. *et al.*, 2002) underlined two main challenges with respect to this perspective:

1. Individuals must be of well-defined and subtle ethnic origin groups. Data collection must be done with a high degree of specificity, along with a suitable definition of ethnicity groups with respect to the trade-off between specificity and reliability.

2. Individuals from mixed-ethnic groups pose greater challenges, thus they must be considered appropriately. The investigator has to rely on multivariate models for adjustment when it is difficult, but not impossible to find matching controls from individuals with multiple ethnic origins.

1.2.2.3 Family-Based Test of Linkage and Association

When matching controls by ethnicity is very difficult or even impossible, the use of family-members as controls can help to overcome this (Thomas, D. C. *et al.*, 2002). The most commonly used familial-based case-control method involves the use of siblings or parents as controls (Schaid, D. J. *et al.*, 1998). One of the major practical difficulties of this is that not all cases will have an available sibling, and yet the sibling control must have already survived free of the disease to the age when the case has been diagnosed. Alternatively, cousins can be used as controls, but again are less readily available for closer matching on age and year of birth. An increasingly used family-member case-control study design uses parents as controls (Thomas, D. C. *et al.*, 2002). A special case of this design is the transmission disequilibrium test (TDT), which tests for systematic differences between the genotypes of diseased children and those expected under Mendelian randomization of alleles of their unaffected parents. TDT allows both linkage and association tests, respectively robust to population structure (causal allele) or subject to fine-scale mapping (non-causal allele in LD in causal allele). While homozygous parents are uninformative, alleles from heterozygous parents are assumed to be independent, implying a multiplicative disease model (Astle, W. *et al.*, 2009).
Let n_A and n_a be the number of A and a alleles transmitted to affected children by Aa heterozygous parents, respectively. Linkage infers that each parental allele is equally likely to be transmitted, thus the null hypothesis for the TDT can be formulated as

$$H_0 : \mathbb{E}\left[n_a\right] = \mathbb{E}\left[n_A\right].$$

Conditional on the number of heterozygote parents $n_a + n_A$, the test statistic n_a has a Binomial$(n_a + n_A, 1/2)$ null distribution. However, the McNemar's test that can approximate a Chi-square with 1 degree of freedom χ^2_1 null distribution is widely used and defined as

$$\frac{(n_a - n_A)^2}{n_a + n_A}, \tag{1.20}$$

The TDT can also be derived from logistic regression models where transmission is the outcome variable, and the parental genotypes are predictors (Schaid, D. J. *et al.*, 1998; Visscher, P.

M., 2010). Though TDT has several advantages, the main disadvantages of the TDT and other above-mentioned family-based case-control methods are the difficulty of obtaining enough families for a well powered study and the additional high costs of this $1:3$ matched case-control study, where three individuals have to be genotyped to obtain the equivalent of one case-control pair (Astle, W. *et al.*, 2009; Visscher, P. M., 2010).

1.2.2.4 Genomic Control

Recently, investigators have proposed the use of genomic information to help address the problem of bias attributable to PS and over-dispersion attributable to CR. If a test statistic T has an asymptotic χ^2_1 distribution under the null hypothesis of no association, it has been demonstrated that in the presence of confounding factors, the test statistic is distributed as χ^2_1 up to some scaling constant (Astle, W. *et al.*, 2009; Devlin, B. *et al.*, 1999). Using a simulation under an island model with admixture and ascertainment bias, Astle and Balding in (Astle, W. *et al.*, 2009) demonstrated an approximately linear inflation of the Armitage test statistic due to a combination of PS and CR, and Devlin and Roeder in (Devlin, B. *et al.*, 1999) argued that this inflation holds more generally. They therefore proposed to correct for type I error from the test statistic T by adjusting all the test statistics by a constant factor λ, keeping the ranking of markers in terms of significance unchanged, that is, GC appears as an adjustment of the significance threshold (Astle, W. *et al.*, 2009). The main idea here is to multiply all the test statistics with λ in order to make the T distribution to fit a χ^2_1 distribution. A number of constants have been proposed to estimate λ, including the *mean estimator* "mean $\left(T^2\right)$ ", the *median estimator* "median$(T^2)/0.455$" and the *regression estimator* "slope of regression of observed T^2 on the expected" (Astle, W. *et al.*, 2009; Devlin, B. *et al.*, 1999).

The *mean estimator* is more effective than *median estimator* under the null hypothesis. However, given that, for most diseases, only a few markers are expected to have strong causal associations with test statistics in the upper tail of the empirical distribution, the portion of empirical distribution that is away from the upper tail can not be used to estimate λ because it should normally reflect the null distribution of association. Astle and Balding in (Astle, W. *et al.*, 2009) proposed to remove the highest test statistics from consideration and then use the mean estimator.

Lemma 1.2.1. *The mean of the smallest* $100q\%$ *values in a large random sample of* χ^2_1 *statistics has expected value*

$$\frac{1}{q} d_3\left(d_1^{-1}(q)\right), \tag{1.21}$$

where d_k *is the distribution function of a* χ^2_1 *random variable.*

Proof. Let X χ_1^2, then

$$\begin{aligned}\mathbb{E}\left(X|X < d_1^{-1}(q)\right) &= \int_0^{d_1^{-1}(q)} x\frac{1}{q\sqrt{2\pi}}\frac{e^{-x/2}}{\sqrt{x}}dx \\ &= \int_0^{d_1^{-1}(q)} \frac{1}{q\sqrt{2\pi}}\sqrt{x}e^{-x/2}dx \\ &= \frac{1}{q}d_3\left(d_1^{-1}(q)\right).\end{aligned}$$

□

Ultimately, Estimate(λ) = mean(lower 95% of T^2)/0.759. A deviation of λ from 1 suggests that the analysis model failed to account for the population structure, and another model should be used. GC assumes an additive disease model while it can be adapted for tests of other disease models. In addition, GC assumes that PS and CR act in the same direction across all loci which is not true. In other words, it does not distinguish markers at which the pattern of association is correlated with the underlying pedigree from those at which the pedigree does not contribute to the association, for which no adjustment should be performed. Finally, it is also important to consider that the inflation factor λ is dependent on sample size, so special methods should be used when the number of people typed for different markers differs (Astle, W. *et al.*, 2009).

1.2.2.5 Structured Association

With the high availability of several thousand genome-wide markers, the ethnic group or subpopulation of an individual in a population can be more confidently inferred. After determining for each individual the ethnic group membership, which is assumed to possess no population structure, any test statistics can be applied within each subpopulation to estimate a correct type I error rate. This process is referred to as the Structural Association (SA) method, based on the island model of population structure, and the ancestry of each individual is assumed to be drawn from one or more islands (Zheng, G. *et al.*, 2012; Astle, W. *et al.*, 2009). The null hypothesis H_0 of no association is tested between a marker and the disease at the subpopulation level, assuming that the subpopulation is in Hardy-Weinberg equilibrium, as in (1.13), rather than the overall null hypothesis H_0 as in (1.14). Some popular software packages, including STRUCTURE (Pritchard, J. K. *et al.*, 2000) and ADMIXMAP (Hoggart, C. J. *et al.*, 2003), which are available free of charge, can be used to infer the subpopulation memberships of individuals. These methods model variation in ancestral subpopulations along a chromosome as a Markov process. Stratified test statistics for association can then be performed to combine association signals across all subpopulations, where in general, logistic regression models of the form (1.19) are used with the admixture proportions of each subpopulation as covariates, making these approaches computationally intensive. The number of subpopulations can be es-

timated from the population data using an optimized measure of model goodness of fit, which significantly increases the computational burden, although, there is usually no satisfactory estimate of the number of subpopulations, as the population model on which these approaches rely (island model) is not well suited for most human populations. Similar to GC, SA methods are only effective using $\sim 10^2$ SNPs (Astle, W. *et al.*, 2009).

1.2.2.6 Principal Component Adjustment

Price in (Price, A. L. *et al.*, 2010) demonstrated that population structure can be controlled using a logistic regression model of the form (1.18) with the use of ~ 100 widely spaced, putatively neutral SNPs as regression covariates. These covariates are informative about the underlying pedigree and are supposed to eliminate its effect in regression-based tests with the locus of interest (Astle, W. *et al.*, 2009; Price, A. L. *et al.*, 2010), and these approaches are computationally faster than SA and more robust to ascertainment bias than GC, allowing one to take advantage of the flexibility of regression methods.
Principal Components (PCs) are often used to summarize high dimensional data without losing much information. PCs can be used to produce maps closely reflecting the environmental and cultural variations in worldwide populations, as well as population migration. They are ideal for summarizing the genetic markers across the genome and proposed for characterizing population differences and controlling for population structure in genetic association studies by integrating PCs of genome-wide SNP genotypes as regression covariates (Yu, J. *et al.*, 2006; Price, A. L. *et al.*, 2006). The idea here is that the genetic background of an individual can be represented using his/her genetic markers, therefore, individuals with similar variations or similar PC values are likely to come from the same subpopulation (Zheng, G. *et al.*, 2012).

Let X be a $n \times L$ matrix of genotypes initially coded as allele counts (0, 1 or 2), where n and L represent the number of individuals in rows and the number of SNPs in columns, respectively. After standardization of genotypes to zero mean and unit variance, the $n \times n$ matrix representing the kinship matrix $\hat{K}$ introduced in (1.2.2.1) is given by

$$\hat{K} = \frac{1}{L} X X^T, \tag{1.22}$$

Given that $\hat{K}$ is symmetric and positive semi-definite, it can be written as an eigenvalue decomposition

$$\hat{K} = v \Lambda v^T, \tag{1.23}$$

where v and Λ is a matrix whose columns are the eigenvectors, or PCs of $\hat{K}$ and their corresponding diagonal matrix (nonnegative) eigenvalues in decreasing order, respectively.
As for SA and regression control, the main idea motivating PC adjustment is that if measures of ancestry, PCs in this case, can partly explain a correlation between the phenotype and a

given SNP, then including these PCs as regression covariates prevents that part of the signal association from contributing to the test statistic. This allows protecting against spurious associations, provided that sufficient PCs are incorporated in the model to consistently explain confounding structure. Generally, in a population, the first two PCs v_1 and v_2, jointly, can accurately predict the proportion of an individual's ancestry arising from each of the subpopulations, and because of case-control ratio variations across the subpopulations, they can also to some extent predict the case-control status (Astle, W. *et al.*, 2009). A popular implementation of PC adjustment, the EIGENSTRAT software, for computational reasons, does not include PCs as logistic regression covariates, but instead performs a linear regression of both phenotypes and genotypes, including the first ten PCs (Price, A. L. *et al.*, 2006). Though it is still unclear how many PCs have to be appropriately included to protect the association test from any eventual inflation, experience has suggested that 2 to 15 PCs are typically sufficient. PC adjustment may fail to protect from population structure effects under more complicated and realistic models of population structure; inflation due to family structure and CR for instance is not guaranteed to be ameliorated by PC adjustment (Astle, W. *et al.*, 2009; Price, A. L. *et al.*, 2010).

1.2.2.7 Mixed Regression Models

Yu et al. in (Yu, J. *et al.*, 2006), as well as many other investigators, have proposed the use of mixed models to model population structure, family structure and CR. The main idea here is to model phenotypes using a mixture of fixed effects, the candidate SNP and optional covariates including age or sex, and random effects, a phenotypic covariance matrix as a sum of heritable and non-heritable random variation (Price, A. L. *et al.*, 2010), Let us consider the standard linear model described in Section 1.2.2.1, extending (1.18) by integrating for each individual $i \in \{1 \dots n\}$ a latent variable δ_i such as

$$\mathbb{E}\left[y_i|\delta_i\right] = \alpha + x_i\beta + \delta_i, \tag{1.24}$$

where δ_i denotes the random effects, which in the context of a case-control study can be interpreted as a polygenic contribution to the phenotype, resulting in many small additive genetic effects distributed across the genome (Price, A. L. *et al.*, 2006), Assuming an additive polygenic, the δ variance-covariance structure that is proportional to the correlation structure of the genotypes, which from (1.15) is proportional to the kinship matrix K, can be expressed as

$$\begin{pmatrix} \delta_i \\ \vdots \\ \delta_n \end{pmatrix} =: \delta \sim \mathcal{N}\left(\mathbf{0}, 2\sigma^2 h^2 K\right), \tag{1.25}$$

and the residuals are assumed to satisfy

$$y_i \left(\alpha + x_i + \delta_i\right) \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \left(1 - h^2\right) I\right), \tag{1.26}$$

where the parameter σ^2 relates K to the phenotype y. By capturing the extent to which genetically similar individuals are phenotypically similar, this enables removing of confounding effects, whereas $h^2 \in [0, 1]$ is the narrow sense heritability of the trait, which represents the proportion of variation due to additive polygenetic effects.

Mixed models have proven useful theoretically but are very computationally extensive. Typically, $\sim 10^5$ SNPs are required for adequate estimation of K in human populations, which is much more than what is required in PC adjustment. However very recent computational advances have now made their application in population-based genetic association possible. The software package EMMAX, developed by Kang et al. in (Kang, H. M. *et al.*, 2010), enables fast inference in linear mixed models using a likelihood ratio test. Another fast method for inference of mixed linear models, TASSEL, proposed by Zhang et al. in (Zhang, Z. *et al.*, 2010), uses a compression approach, called 'compressed MLM', that decreases the effective sample size of large datasets by clustering individuals into groups, markedly reducing computing time while maintaining or improving statistical power.

1.3 Genome-wide association studies: progress and limitations

The ultimate goal of genetic association studies is to identify genetic risk factors, and elucidate the mechanisms through which these factors exert their effects on rare Mendelian diseases such as sickle cell anemia and cystic fibrosis (Bush, W. S. *et al.*, 2012; Wright, F A. *et al.*, 2011), and on common complex diseases such as heart disease (HD), type 2 diabetes (T2D) and tuberculosis (TB). This goal, is not only limited to identifying the genetic factors, it extends also to providing an overall genetic architecture including estimating the genetic heritability, the number of loci underlying phenotypic variation, and the distribution of effect sizes, as well as suggesting whether or not the effect of epistasis or pleiotropy exist (Stranger, B. E. *et al.*, 2011; Hardy, J. *et al.*, 2009).

1.3.1 Overview of GWA studies

Candidate gene association studies interrogate human genetic variations for association with complex diseases, prioritizing certain genes or genomic regions for further investigation. Recently, genetic association studies have utilized LD-based (1.1.3) genome-wide association (GWA) studies, which permit interrogation of allele frequencies at each of several hundred thousand markers spaced throughout the entire human genome at levels of resolution previ-

ously unachievable, in thousands of unrelated case and control individuals. GWA studies are unbiased with respect to genomic structure and prior hypotheses regarding genetic association with the disease, in contrast to candidate gene studies, where the knowledge of the trait is used to identify candidate loci involved in the etiology of the disease of interest (Pearson, T. A. *et al.*, 2008). GWA studies have also been valuable over family-based linkage studies (see 1.2.2.3), which, though successful in identifying genes of large effects in Mendelian disease, have had less success in common diseases that are in most cases the results of multiple genes generally of small effects (McCarthy, M. I. *et al.*, 2008; Bush, W. S. *et al.*, 2012).

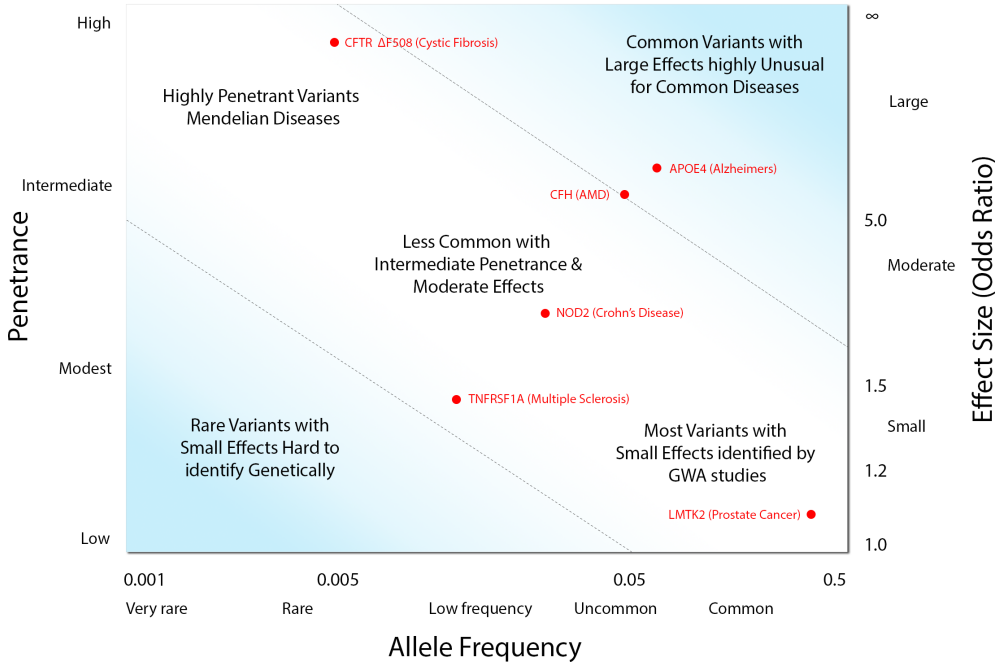

Figure 1.8: The spectrum of disease association can be conceptualized in a 2-dimension representation as function of allele frequency and penetrance or effect size. The bulk of identified genetic associations lie on the diagonal denoted by the dashed lines. Highly penetrant variants with large effect sizes usually for Mendelian diseases are extremely rare, while common variants that are mostly identified through GWA studies are of small effect sizes and have very low homozygote risk penetrance.

Though unbiased with respect to previous biological knowledge, and with respect to the genome location, GWA studies are not unbiased with respect to what is detectable (Visscher, P. M. *et al.*, 2012). In fact, GWA studies rely on the *common disease, common variant* (CDCV) hypothesis, suggesting that genetic factors influencing many common diseases will be at least to some extent attributable to a limited number of allelic variants present in more than 1% to 5% of the population. Therefore, as illustrated in Figure 1.8, GWA studies are by design underpowered to detect association with variants that are relatively common with modest effect sizes (OR of 1.1 to 1.5), even when combined, their overall impact on the population variance and predictive power still remains limited, and many important complex disease-causal variants

may be rare, and are therefore unlikely to be identified through this approach (Stranger, B. E. *et al.*, 2011; Visscher, P. M. *et al.*, 2012; Pearson, T. A. *et al.*, 2008).

1.3.1.1 Progress of GWA studies

The first successful GWA study was reported in 2005 with $\sim 100,000$ SNPs from 96 age-related macular degeneration patients and 50 healthy controls (Klein, R. J. *et al.*, 2005), and since then several hundred impressive GWA studies have been published covering a range of complex human traits. A landmark study by the Wellcome Trust Case Control Consortium (WTCCC) in 2007 was carried out on seven different common diseases including bipolar disorder (BP), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthrisis (RA), type I diabetes (T1D), and type 2 diabetes (T2D), covering $\sim 500,000$ SNPs from $1500 - 2000$ cases and 3000 shared controls (Wellcome Trust Case Control Consortium, 2007). As of December 2013, $1,779$ GWA studies and $12,126$ genetic variants have been published according to The Catalogue of published Genome-Wide Association Studies. These cover a broad of spectrum of complex traits and diseases conducted in an unprecedented global research effort in only 8 years, including hundreds or even thousands of case and control participants. Through the experimental design of GWA studies, previously implicated loci, and also a number of novel associated loci for complex diseases, have been discovered to be significantly associated with one or more complex traits (Hindorff, L. A. *et al.*, 2013b), including autoimmune diseases such as type 1 diabetes (T1D), ulcerative colitis (UC) (Anderson, C. A. *et al.*, 2011) and Crohn's disease (CD) (Rioux, J. D. *et al.*, 2007; Mathew, C. G., 2008; Franke, A. *et al.*, 2010); metabolic diseases such as type 2 diabetes (T2D) (Saxena, R. *et al.*, 2007), fasting glucose and insulin levels, body-mass index (BMI) and obesity (McCarthy, M. I., 2010).

1.3.1.2 Study designs used in GWA studies

The most frequently and substantially used design for GWA studies to date has been the case-control design (see 1.1.1). These studies are often easier to conduct and cost effective compared to other designs, particularly in cases of sufficient numbers of case-control individuals or sample size. Additional assumptions are made in this design, which, if not met, can lead to substantial biases and spurious associations (see 1.2.2). That said, the statistical power of GWA studies depends on certain factors including phenotype definition, sample size, effect size, causal allele frequency, and marker allele frequency as well as its relationship with the causal variant (Platt, A. *et al.*, 2010; Pearson, T. A. *et al.*, 2008; Spencer, C. C. *et al.*, 2009). Therefore, it is required for instance, to have large population samples to be able to detect variants of even moderate effect sizes (OR of 1.5 to 2.0). Statistical power and sample size can be substantially increased through *meta-analyses* of independent GWA studies for the trait of interest. Several software packages are available for carrying out meta-analysis such as METAL (Willer, C. J. *et al.*, 2010). When different array technologies and different marker coverage are used, only a few

SNPs may be common, therefore, genome-wide imputation methods, recently developed and implemented in software packages such as IMPUTE and MaCH, can be of great benefit to infer genotypes of untyped SNPs using a reference panel of more densely genotyped samples (e.g. The International Hapmap Project - http://hapmap.ncbi.nlm.nih.gov/, and the 1000 Genomes project - http://www.1000genomes.org/) (Stranger, B. E. *et al.*, 2011; Howie, B. N. *et al.*, 2009; Hancock, D. B. *et al.*, 2012).

Because the hundreds of thousands to millions of single-locus statistical tests conducted in GWA studies, expressed as $p-$values, each having a false positive probability relative to its test statistics, the cumulative likelihood of detecting one or more false positives among the whole GWA analysis is much higher. Simple ways to correct for multiple testing approaches include the *Bonferroni correction*, and the *false discovery rate* (FDR). Another approach for establishing significance in GWA studies is using *permutation testing*, although a bit more computationally expensive, this approach is a straightforward way to generate empirical distribution of test statistics of a data set assuming the null hypothesis is true (Bush, W. S. *et al.*, 2012). In addition to single-locus testing, contemporary association studies consider multi-locus analysis, whereby interactions among genetic variants across the entire genome are examined. While presenting enormous opportunity to examine multiple, often correlated phenotypes or multiple genetic models, this approach still presents various statistical and computational challenges, and thus it is not as straightforward as a single-locus analysis (Lunetta, K. L., 2008).

1.3.1.3 Quality control in GWA studies

The selection of study participants remains of key importance in a GWA study, as no genetic association study will have meaningful outcomes without a precise and consistent characterization of the phenotype of interest. Misclassification of participants, cases or controls, can remarkably reduce study power, and introduce systematic biases leading to an increased number of false-positive and false-negative association signals. Genotyping errors, especially when occurring differentially between case and control individuals, can also lead to spurious associations. Although such errors can be avoided through study design (see also 1.2.2.2), a thorough data quality assessment should be applied both on a per-sample and a per-marker level to remove individuals or markers with particularly high error rates (Pearson, T. A. *et al.*, 2008; Anderson, C. A. *et al.*, 2010).

Checks on samples consist of at least 4 steps including (1) identification of samples with inconsistent sex information; (2) identification of samples with outlying missing genotype or heterozygosity rate; (3) identification of related or duplicated samples and (4) identification of samples of divergent ancestries. For example, as DNA sample quality or concentration can have large effects on genotype call rate and genotype accuracy, samples with low DNA quality

must be discarded. The genotype failure rate and heterozygosity rate can be used to measure DNA quality and only samples with less than $3-7\%$ missing genotypes are typically reported and considered for the further analyses. In this case, the most appropriate threshold, which depends on the sample, can be determined by carefully scrutinizing the genotype missing rates across the entire data set. Duplicated and related individuals can be identified using the IBS and IBD metrics. After calculations of IBS and IBD between all pairs of individuals, duplicated individuals have an IBS $= 1$ or IBD $= 1$, however, due to genotyping error, PD and population structure, which can often cause some variations to these theoretical values, an IBD > 0.98 denotes duplicates, and one of each pair of individuals with IBD > 0.1875 must typically be removed.

Once samples failing per-individual quality control (QC) are removed, individual SNPs across the remaining samples undergo further assessments for probable genotyping errors (per-SNP QC), including (1) SNPs with an excessive missing genotype; (2) SNPs with severe deviations from the Hardy-Weinberg equilibrium; (3) SNPs with excessive differences in missing genotype rates between cases and controls and (4) SNPs with very low minor allele frequency, as rarer SNPs are difficult to measure reliably in case-control design (see Figure 1.8). There exist several software packages for QC of GWA data including GenABEL (Aulchenko, Y. S. *et al.*, 2007), snpMatrix (an R package part of the bioconductor project - http://www.bioconductor.org) and Plink (Purcell, S. *et al.*, 2007), which have advantages over standard statistical software. Currently, Plink is the most widely used software package for GWA data QC, using the QC protocol presented by Carl et al. in (Anderson, C. A. *et al.*, 2010), with fully automated analyses.

1.3.2 Limitations and future directions of GWA studies

GWA studies promise to provide an understanding of the mechanisms underlining the etiology and pathogenesis of complex diseases. A fact sheet published by the National Human Genome Research Institute (NHGR) in 2007, states that the impact of GWA studies on clinical medicine could potentially be substantial and that this could ultimately lead to the era of personalized medicine, enabling more customized clinical strategies, including individual risk prediction, disease prevention, and patient-specific treatment (Hindorff, L. A. *et al.*, 2013a). From a medical genetics perspective, the ultimate goal of GWA studies can simply be viewed as the identification of genetic risk factors from the validated disease-SNP associations, in order to characterize their functional effects on the disease. When many genes are believed or suspected to be involved in the pathogenesis of a disease, and when the understanding of the etiology of the disease, a systems biology perspective should be taken into consideration, in which perturbations of complex networks are considered to be the basis for the outcome of a complex trait phenotype (Stranger, B. E. *et al.*, 2011). GWA studies have provided important

scientific discoveries as discussed in Section 1.3.1.1 and have had an enormous impact on the field of human genetics, but at the same time, many people have highlighted several problems and limitations of this single-marker-based experimental design. Those limitations include not only, but importantly, methodology issues such as insensitivity to rare and structural variants, requirement of large sample sizes, eventual biases due to selection of study participants, and genotyping errors. Other limitations include research outcome issues such as the potential for spurious results, utility of the research findings such as lack of information on gene function, heritability remaining unexplained in the population, limited information about non-genetic risk factors, and lack of biological and clinically relevant or meaningful results (Visscher, P. M. *et al.*, 2012; Pearson, T. A. *et al.*, 2008). These limitations mean, the single-marker-based approach of GWA studies might not possess adequate power to detect important risk factors for complex disease because:

1. The majority published GWA studies enumerate only the $20-50$ most significant SNPs and their neighboring genes (the "most-significant" SNPs/genes approach), while paying little or even no attention to the rest, thus neglecting most genetic variants with small effects on the disease. This approach might even miss genetic variants with higher disease risks, as those variants might not rank among the top $20-50$ most-significant out of hundreds of thousands of markers tested, especially in cases where the size of the sample is not large enough (Wang, K. *et al.*, 2007).

2. This strategy has limited power to unveil the majority of genetic disease risk factors. For instance, these have been estimated to account for about 80% of the heritability for the human height trait (Visscher, P. M., 2008), however the total of phenotypic variance explained by more than 40 variants discovered so far in human height GWA studies, only sums to about $\sim 14\%$. Most of the complex disease associated variants identified to date together account for much less of the phenotypic variance (Stranger, B. E. *et al.*, 2011).

With the exception of age-related macular degeneration (AMD) (Klein, R. J. *et al.*, 2005) and type 1 diabetes (T1D) (Barrett, J. C. *et al.*, 2009), for which collectively the proportion of heritability explained to date is approximately 50% and 80%, respectively, most complex disease variants identified to date, using GWA study design, together account for much less of the trait variance. Several proposed explanations for this "missing heritability" (Stranger, B. E. *et al.*, 2011; Manolio, T. A. *et al.*, 2009) include:

1. Effect sizes of associated variants may be underestimated due to incomplete linkage disequilibrium between causal variants and marker SNPs;

2. low-frequency polymorphisms (MAF $0.005-0.05$) or rare variants (MAF < 0.005) that are not captured by current genotyping platforms, including Copy Number Variants (CNVs), may contribute a portion of the unexplained heritability;

3. heritability may be overestimated, with epistasis, epigenetics, and genotype–environment interactions contributing to trait heritability; and

4. many additional, currently undetected small effects may together comprise a significant contribution to heritability.

Structural variants such as insertions/deletions (indels), inversions and CNVs, as well as rare variant have been made available from various human genome projects. Along with the launch of the 1,000 Genomes Project in 2008, sequencing the genomes of over 2,000 individuals and providing low-frequency variants in the human genome, a next wave of GWA studies has been facilitated to investigate the role of genetic variants not yet explored. This may enable the discovery of genetic risk variants of moderate and high effect sizes, to cover the substantial heritability that remains unexplained. In addition, whole-exome sequencing, candidate locus resequencing, and even whole-genome sequencing will also play a key role in the next generation of GWA studies. Integrating this with data from the environment, and other high-throughput technologies such as the proteome and the transciptome, has potential to facilitate further understanding of biological processes, gene-gene and gene-environment interactions, as well as epigenetic effects influencing complex human diseases (Klein, C. *et al.*, 2012).

1.4 Overview on biological networks and pathway databases

The remarkable success of genome-wide association studies has been demonstrated in the uncovering of multiple genetic variants associated with complex diseases, and have clarified our understanding of the genetic architecture of several traits. In a typical GWA study, the single-marker-based approach, hundreds of thousands of markers are simultaneously tested, and the allelic frequencies of each marker between cases and controls compared and evaluated at a genome-wide significance cutoff p-value of 5×10^{-8}, under the assumption of no-association among markers. Because of the very large number of SNP markers used in such studies, only a few markers exceed the genome-wide significance threshold, and in almost all published GWA studies, only the $20-50$ most significant markers are listed (the "most-significant SNPs/genes" approach), while the rest that do not pass this stringent cutoff are generally neglected (see Section 1.3.2). However, many of the markers showing modest associations ($0.05 >= p > 10^{-08}$), though with small but measurable genetic effects, may represent false negatives, in which case, accepting the null hypothesis of no-association represents a type II error (Baranzini, S. E. *et al.*, 2009). In addition, the very small fraction of heritability explained by identified genetic variants in most GWA studies so far (see Section 1.3.2), suggests that there is a substantial proportion of risk alleles or genetic variants that are still missing, and can not be detected through this most-significant strategy (International Multiple Sclerosis Genetics Consortium, 2013).

Realizing the limitations of conventional single-marker association analyses, alternative or complementary approaches for GWA study analysis have been developed in recent years. These include association tests that use multiple SNP markers, association tests using imputed genotypes, those incorporating linkage information and, more recently, pathway-based analysis (PA) approaches. One of the first studies to propose a PA approach for GWA study analysis was conducted by Wang et al. in (Wang, K. *et al.*, 2010) where, borrowing ideas from the gene expression microarray field, they proposed an adapted GSEA (Gene Set Enrichment Analysis) approach (Subramanian, A. *et al.*, 2005) to use pathway information in GWA studies, and demonstrated that this approach might complement the most-significant SNPs/genes approach for interpreting GWA results on complex diseases. The main idea behind PA in GWA studies is that, although association signals of several variants involved in disease etiology may be too small to detect using the conventional single-marker-based approach, they may be collectively detected from the combined effect of multiple variants in interacting or grouped genes according to their shared functions. Therefore, integrating prior biological knowledge on genes and pathways into association studies, may increase the power to identify not only genes, but also mechanisms that influence the susceptibility and development of complex diseases (Wang, K. *et al.*, 2010; Wang, K. *et al.*, 2007; Fehringer, G. *et al.*, 2012; Baranzini, S. E. *et al.*, 2009; International Multiple Sclerosis Genetics Consortium, 2013).

Identification of biological functions of disease-related genes that are as yet unknown, will substantially increase our understanding of biological mechanisms involved in disease pathogenesis. As it is more likely that proteins that are closely connected collaborate in similar processes, understanding functional roles of interacting partners may facilitate the understanding of potential functional roles of as yet unannotated disease-related genes, and will increase our understanding of biological processes underlining complex diseases. These biological processes are generally represented as networks interconnecting a very large number of nodes and are characterized by very complex topologies (Ma'ayan, A., 2011). Several biological networks have been characterized to date including transcriptional regulation networks (Wei, C. L. *et al.*, 2006), protein-protein interaction networks (Keshava Prasad, T. S. *et al.*, 2009), metabolic networks (Karp, P. D. *et al.*, 1998), cell signaling networks (Ma'ayan, A. *et al.*, 2005), epistasis interaction networks, in which genes are connected only if they exhibit a genetic interaction when knocked out or down-regulated (Segre, D. *et al.*, 2005), disease gene interaction networks in which diseases are connected to genes that, when mutated, contribute to the disease (Goh, K. I. *et al.*, 2007), and drug interaction networks, in which drugs are connected to their targets (Yildirim, M. A. *et al.*, 2007).

Mathematical objects such as graphs, consisting of nodes and edges interconnecting these nodes, can offer a convenient and effective level of abstraction for representing the aforementioned complex networks. Widely used to represent mathematical networks, different compu-

tational approaches, such as data mining, machine learning, search or optimization problems and statistical approaches, can be designed based on this graphic representation to reveal the organization of these complex networks at different levels. Biological scientists too can benefit from graph-based formalism since they are able to visualize the overall topology of the network, modify nodes and links and annotate them with additional information. The robustness and evolution mechanism of a biological network, for instance a cell signaling pathway, is believed to be dependent on its topological structure (Strangio, M. A., 2009).

1.4.1 Analysis of biological networks

Complex systems such as social organizations and biological systems can be represented as networks wherein components of the systems are represented as nodes and are linked by edges representing their interactions or simply their relationships. In the case of biological networks, components such as proteins, genes, diseases and drugs are typically considered as nodes and their direct or indirect interactions as edges (A.L., B. *et al.*, 1999). The directionality of such networks depends on the biological characteristic components, for example, protein-protein and genetic interactions are usually represented with undirected networks, and transcription factor binding and metabolic networks are represented as directed networks as their components have certain directions in their interactions (A.L., B. *et al.*, 1999). The complexity of biological networks is not only driven by their large number of nodes and their interactions, but also by other complex dynamic behaviors difficult to observe in the dynamic cellular 3D-space. In protein-protein interaction networks for instance, interactions are often subject to time and location, and proteins may vary their partners accordingly. In this case, quantitative mathematical methods may be computationally extremely expensive or even unfeasible. Therefore, representing the complexity of biological systems as networks may have several advantages such as for structural analysis of these biological systems, as well as in hypothesis formulation. This could substantially increase insights into the organizational principles of biological components and reveal information on the underlying biological functions (Strangio, M. A., 2009; Zhu, X. *et al.*, 2007). Once represented by a network, a biological process or system can be studied using tools of complex network theory, to systematically characterize its functional properties. Network topology is of crucial importance in the architecture and robustness of a network, this includes information about the global topological properties of the entire network, general and specific properties of nodes or edges, and modules within the network (Ma'ayan, A., 2011).

Let G be the mathematical representation of a biological network, defined as a couple (V, E) and composed of a set of nodes or vertices V and edges E, where $E = \{(i, j) \,|\, i, j \in V\}$. We define by $G' = (V', E')$ a *subgraph*, where $V' \subseteq V$ and $E' \subseteq E$ and each edge in E' is incident with nodes in V'. A *clique* in an undirected graph G is a subgraph G', in which every pair of nodes is adjacent. The *degree* of a node is the number of edges incident to the node. In

an undirected graph, the degree of a node i is defined as $deg\,(i) = k\,(i) = |N\,(i)\,|$ with N, the number of neighbors of node i. The *degree distribution* $P\,(k)$, describes the probability of a node to be of degree k. The *distance* $\delta\,(i,j)$ is the smallest number of links that have to be traversed to get from node i to node j, and the path through the network that achieves that distance is the *shortest path* between nodes i and j. In an interaction network, the *graph diameter* and the *average of shortest path lengths* l refer to the maximum and average value of δ between any two nodes $i, j \in V$, with $\delta \geq 1$. The *clustering coefficient* of a node i in a network, C_i, is the probability that two given nodes i and j which are connected to the node t are themselves connected. In an undirected graph G is given by $C_i = \frac{2E_i}{k_i(k_i-1)}$, where k is the degree of node i, and E is the number of edges between the k neighbours of i in G, with $0 \leq C_i \leq 1$. The tendency of the whole network to form clusters can be measured using the *average clustering coefficient*, and is defined as $\overline{C} = \frac{1}{N}\sum_{i=1}^{N} C_i$, where $N = |V|$ is the number of nodes.

It is generally assumed that there are small values of l and $\overline{C}$, and small l implies that a given node i can reach any other node j in a few steps. Thus this acts as a measure of the information efficiency flow across the network, whereas a small $\overline{C}$ implies that the network is likely to not contain clusters (Albert, R. *et al.*, 2002). Several recent studies, including (Albert et al., 2002), have revealed that many complex systems present *small-world* and *scale-free* global network topological properties. Small-world networks are characterized by significantly larger $\overline{C}$ and relatively small l. On the other hand, scale-free networks exhibit $P\,(k)$ that follows the power-law distribution $P\,(k) \sim k^{-\gamma}$, indicating that the fraction $P\,(k)$ of nodes in the network with degree k is proportional to $k^{-\gamma}$ where γ denotes the *degree exponent*, typically in the range $]2, 3[$. In contrast to *random networks*, where the process of adding nodes is purely probabilistic with all the nodes having equal importance, in scale-free networks, the process of node addition follows a preferential rule, meaning that when a new node is added to the network, it is attached preferentially to the nodes with the highest degrees, leading to the power-law degree distribution (Strangio, M. A., 2009; Albert, R. *et al.*, 2002). The latter topology is likely to be the most reasonable and desirable for several biological networks (including protein-protein interactions, transcription factor binding, metabolic, and genetic networks) given that it converges toward a structure that enables efficiency of its functions (Strangio, M. A., 2009). This can be explained by the evolutionary character of biological networks, where nodes with high degrees, important for the survival of the organism, have resisted selective pressure. Moreover, this topology makes such networks tolerant to random failures and errors such as missense point mutations, but more vulnerable to targeted attacks, as the removal of nodes with the highest degrees is highly disruptive for the networks (Albert, R. *et al.*, 2002).

Properties of nodes in a network are important, for instance in detecting central or intermediate

nodes that affect the topology of the network. Depending on the question asked, these properties can be crucial in biological networks, for instance, in finding nodes that are not necessarily central or possessing a high number of links ("hubs"), but that have a critical biological role in PPI networks. These properties include most basically, connectivity degree (k), or a centrality measure that characterizes each node or edge according to their contextual location within the network such as the closeness centrality, the eigenvector centrality, betweenness centrality, eccentricity centrality and subgraph centrality. Properties of edges include edge betweenness centrality, the types of relationship that represent inhibiting or activating relationships between a pair of nodes, and edge directionality, which indicates that the upstream and downstream nodes are connected by a specific link (Ma'ayan, A., 2011; Pavlopoulos, G. A. *et al.*, 2011). Centrality measures are discussed in more detail in Chapter 2.

Although the global network topological and node-edge-specific properties have been extensively used to analyse the topological structure of biological networks, recently much attention has been paid to the local units of the networks, called network modules or network clusters, which represent dense parts of connectivity demarcated by regions of low connectivity (Ma'ayan, A., 2011; Zhu, X. *et al.*, 2007). Several methods have been developed to find possible modules in networks, including the traditional method for hierarchical clustering, unsupervised clustering algorithms, such as nearest neighbor clustering, Markov clustering, and betweenness centrality-based clustering, which uses nodes with high betweenness centrality and low connectivity to separate clusters (Newman, M. E., 2001).

Despite the fact that network modules are still ubiquitous structures in most biological networks, understanding the topological structure of such networks, using topology analysis measures mentioned above and others, such as network motifs, may help to better characterize and understand the interplay between network structure and function (Ma'ayan, A., 2011).

1.4.2 Protein-protein interaction databases

The majority of biological processes in a living cell are driven and mediated by proteins, which transport or store other molecules, act as catalysts, confer immunity, transmit signals, and control growth and development. Most proteins function by interacting with other molecules including lipids, nucleic acids, or other proteins. Proteins do not operate independently, rather in close collaboration with partner proteins or in complex protein networks, to carry out complex biological activities. PPI networks exhibit scale-free network properties, and one advantage of scale-free networks is that the robustness-loss of individual components generally maintains overall network topology, making the system relatively immune to defects in individually targeted components (Zhu, X. *et al.*, 2007). Many forms of cancer for example, require the loss of multiple components for network breakdown, which may explain, in part, the observation that processes leading to the onset of cancer are often the results of multiple mutations (Knudson,

A. G., Jr., 1971). Nevertheless, some regions of networks should be more vulnerable to network breakdown than others, that is, mutations that exert an influence on hubs and other important network components are more likely to cause system defects than those affecting the periphery (Zhu, X. *et al.*, 2007).

Protein-protein interactions (PPI) are generally generated by experimental methods, such as coimmunoprecipitation, the yeast two-hybrid system or affinity purification, genomic context or phylogenetic profiling approaches, protein microarrays or synthetic lethality (Bjorn, H. J. *et al.*, 2008; Lehne, B. *et al.*, 2009; Hart, G. T. *et al.*, 2006). PPI networks represent some of the largest and most diverse datasets available to date, and, to be effectively exploited, they need to be stored in a consistent and reliable way (Zhu, X. *et al.*, 2007). Recently, several PPI databases have been developed, collecting published PPI data, and providing curated datasets for public access (Lehne, B. *et al.*, 2009). However, due to different methods used for data collection, as well as different data formats in which these datasets are represented, these PPI databases differ considerably in their coverage and content. The difference in methods mentioned above also cause some confusion with respect to the meaning of the term "protein-protein interaction", which could be direct physical binding, membership of the same multi-protein complex, or a functional interaction (Hart, G. T. *et al.*, 2006).

Table 1.5: Protein-protein interaction (PPI) databases.

Database	URL	Proteins	Interactions	Publications	Organisms
HPRD	http://www.hprd.org	30,047	41,327	453,521	1
BioGRID	http://www.thebiogrid.org	54,549	500,239	42,172	49
MINT	http://mint.bio.uniroma2.it/mint	35,553	241,458	5,554	144
IntAct	http://www.ebi.ac.uk/intact	81,209	439,013	12,397	131
DIP	http://dip.doe-mbi.ucla.edu	26,453	76,8441	6,678	649

More comprehensive descriptions of PPI databases have been reported in excellent recent reviews (Lehne, B. *et al.*, 2009; Rohl, C. *et al.*, 2006) and the complete listing is available at http://www.pathguide.org (Bader, G. D. *et al.*, 2006). Table 1.5 lists some major public databases, on which we focus in this work, which combine protein interaction data from both experimental methods mentioned above and data mining of the literature. These databases include the Human Protein Reference Database (HPRD) (Peri, S. *et al.*, 2004), the Biological General Repository for Interaction Datasets (BioGRID) (Stark, C. *et al.*, 2006), the Molecular INTeraction database (MINT) (Chatr-aryamontri, A. *et al.*, 2007), the IntAct molecular interaction database (IntAct) (Chatr-aryamontri, A. *et al.*, 2007), and the Database of Interacting Proteins (DIP) (Xenarios, I., 2002). The information in Table 1.5 is based on datasets downloaded from the individual databases in January 2014, containing complete sets of binary

interactions. Only IntAct and MINT provide information about the model applied to derive the interactions, "the spokes model", while no other database provides this information. Though most databases also provide *genetic interactions* which refer to the interaction between two non-essential genes that lead to a non-viable phenotype if they are knocked out simultaneously (Lehne, B. *et al.*, 2009), only *physical interactions* are considered in the subsequent analysis.

Currently, BioGRID is the most comprehensive database in terms of unique interactions, with more than 500,239 from 49 different organisms, citing up to 42,172 different publications. We denote considerable increases in the BioGRID database not only in terms of unique interactions, as well as in the number of organisms, but also with respect to different publications cited, expanding from 90,972 to 500,239, 10 to 49, and 16,369 to 42,172, respectively, from 2008 to 2014 as reported in (Lehne, B. *et al.*, 2009). Surprisingly, although being restricted to human proteins, HPRD cites over 453,521 publications for 41,327 interactions for only one organism. However, it should be noted that the databases examine publications at different depths, with different confidence sets or thresholds or differences in the application of the matrix or spokes model, therefore a higher number of publications should not necessarily be interpreted as a greater curation effort (Lehne, B. *et al.*, 2009).

The overlap between databases is very small (See Table 1.6), making it difficult to obtain confidence in the interactions. Of the 14,899 publications shared by two or more databases, 39% (5,782) were reported with a different number of interactions for different databases in (Lehne, B. *et al.*, 2009), although ideally, every database should extract the same interactions from a given publication. For example, the publication with pubmed ID '14605208' reports 20,405 in its abstract (Lehne, B. *et al.*, 2009), whereas the number of interactions reported by all the five databases citing this publication differ, with the minimum of 20,043 in BioGRID and a maximum of 20,670 interactions in IntAct.

Table 1.6: Overlap of human PPIs between five major databases. The pairwise relative overlap between databases was calculated for *Homo sapiens*, and the absolute numbers are normalized to the total number of PPIs for every row.

	PPIs Total	DIP	MINT	IntAct	BioGRID	HPRD
DIP	3,299		16%	17%	32%	42%
MINT	22,847	2%		31%	25%	38%
IntAct	71,315	1%	10%		12%	7%
BioGRID	138,709	1%	4%	6%		9%
HPRD	38,106	4%	23%	13%	33%	
PPI relative overlap		$0-14\%$	$15-29\%$	$>30\%$		

To assess these difference further, we performed a comparison and estimated the relative pair-

wise overlap only for human PPIs between databases (Table 1.6). HPRD, which focuses on human interactions and reports the most PPIs, has the highest relative overlaps when compared with all other databases. BioGRID, which currently provides the highest number of interactions for human, also presents relatively high overlaps with other databases. Thus, significant overlaps ($> 30\%$) were observed between BioGRID and HPRD, DIP and BioGRID, DIP and HPRD, MINT and HPRD, and between MINT and IntAct.

Though databases analyzed here have improved greatly over the past few years, and addressed some of the important issues, such as data compatibility and exchange. The pairwise overlap between databases performed here doesn't even reach 50%, in contrast with previous studies where up to 75% of pairwise overlaps were reported (Lehne, B. *et al.*, 2009). None of the existing databases provides a fully comprehensive dataset, and as mentioned above, these databases differ not only in the different ways to capture and curate information but also in the ways they represent the data, making the integration of data from the different databases difficult. To overcome this issue, the International Molecular Exchange (IMEX; www.imexconsortium.org) consortium was formed. The main aim of this consurtium is to enable the exchange of data and avoid the duplication of the curation effort, by providing an XML-based proteomics standard, the proteomics standards initiative, for molecular interaction (PSI-MI), as a standard data representation format.

1.4.3 Pathway annotation databases

Current pathway-based approaches for analysis of GWAS data sets may receive as input individual SNP genotypes or a list of $p-$values relating SNPs or genes to the phenotype under study, but can broadly be classified according to the strategy discussed in Subsection 1.5.2. Independently of the overall strategy, these methods utilize functional information on existing annotated pathways. This is vital for the analysis itself for some approaches (i.e. data mining and gene set enrichment analysis methods), and for the interpretation of results for other approaches (i.e. network-based methods). There exist several prominent pathway annotation databases providing diverse features (see Table 1.7), ranging from freeware databases that provide easily accessible and transparent contents, to commercial databases that provide user-friendly statistical software along with high-quality visualization interfaces (Fridley, B. L. *et al.*, 2011). Similar to PPI databases databases (Subsection 1.4.2), most pathway databases rely on manual curation by experts, or alternatively on electronic curation using text-searching algorithms to infer relationships (Ramanan, V. K. *et al.*, 2012). In either curation case, investigators should consider, in accordance with their resources and study goals, criteria used as evidence for inclusion in pathways, text-searching algorithm accuracy, as well as the biological coverage of pathway annotations. Although they supply different features such as transcription

factor networks, gene regulatory networks, signaling pathways, pathway diagrams or metabolic pathways, across these databases, similarly named pathways can present vast differences in their constituting components, whereas significant overlaps can be observed among differently named pathways (Ramanan, V. K. *et al.*, 2012; Fridley, B. L. *et al.*, 2011). This explains some of the differences in results that can be generated from the same input dataset, so careful considerations about which pathway annotation data to be used is recommended, and for a given analysis, multiple databases should preferably be used.

Table 1.7: Some pathway annotation databases.

Name	CM	Description	Cost	URL
BioCarta	M	Driven by user input with expert review of some pathways	F	www.biocarta.com
KEGG	M	Large database of metabolic with reference pathways and organism-specific annotations	F	www.genome.jp/kegg/
PharmGKB	M	Pharmacogenetics and Pharmacogenomics Knowledge Base with a number of drug action pathways	F	http://www.pharmgkb.org/
Ingenuity	M/E	Large collection of canonical pathways from millions of individually; high-quality pathway maps	C	http://www.ingenuity.com/
Reactome	M	Curated database of human biological pathways; A suite of data analysis tools to support pathway-based analysis of experimental data.	F	http://www.reactome.org
Malaria	M	Database for Malaria biology, biochemistry and physiology. Contains numerous metabolic pathways	F	www.sites.huji.ac.il/malaria/
MetaCyc	M	Pathways from over 2391 different organisms; Describes metabolic pathways, reactions, and substrate compounds	FA	http://metacyc.org/
MetaCore	M	Integrates data from different levels of cellular function, including membrane receptors, signal transduction, transcription factors, and effector networks.	C	thomsonreuters.com/metacore/

Abbreviation: CM, Curation method; M, Manual; E, Electronic; F, Free; C, Commercial; FA, Free for academic only

Because of the broad interest in pathway-based analysis for GWA studies, some refinements of pathway annotation databases would be of great interest and may expand their use for investigators from a variety of backgrounds. These enhancements may include user-friendly search and download mechanisms, consistency in pathway naming and classifications, as well as methods for describing overlap between pathways. In addition, a universal file annotation format enabling cross-exchange among analytical tools, might allow investigators more flexibility in precisely matching their databases and statistical methods (Wang, K. *et al.*, 2010; Ramanan,

V. K. *et al.*, 2012).

1.5 Pathway-based approaches for GWAS analysis

The limitations of the conventional single-marker-based approach for GWA studies have urged investigators to explore and develop alternative or complementary approaches for GWA studies (see Section 1.3.2), including, more recently, pathway-based analysis (PA) approaches. In PA of GWA studies, predefined sets of interconnected genes are examined based on prior biological knowledge, and these sets of related genes or pathways are evaluated with respect to their significance based on the disease association signal of markers in or near genes that are part of these pathways (Wang, K. *et al.*, 2010).

1.5.1 Linking pathways to complex diseases

It is now well known that genes do not operate in isolation. They collaborate in groups to carry out specific biological functions, the disruption of which may be related to human diseases. Therefore, depending on the genetic architecture of a complex disease, it is possible that many SNPs or genes, randomly distributed with respect to biological function, having low or moderate risk may interact to confer a significant combined effect for the complex disease (Jia, P. *et al.*, 2011). Schadt, 2009 has recently looked at how networks act as sensors and drivers of common human diseases, and demonstrated, as shown in Figure 1.9, that in order to understand the behaviour of any gene in the context of human disease, individual genes must be perceived in the context of biological networks that define the disease states (Schadt, E. E., 2009).

Many other studies have also investigated and demonstrated the role of complex molecular networks and cellular pathways in the pathogenesis of complex diseases such the autoimmune disease rheumatoid arthritis (RA)(Martinez, A. *et al.*, 2006). For instance, a highly significant interaction was detected between genes associated with the infectious disease tuberculosis (TB) (Hu, P. *et al.*, 2011). In other examples, genes of the complement pathway are suspected to be involved in the disease pathogenesis of age-related macular degeneration (AMD) (Shahbaba, B. *et al.*, 2012), the autophagy pathway was suspected to be implicated in the pathogenesis of inflammatory Crohn's disease (CD) (Shahbaba, B. *et al.*, 2012), and finally the PGE2 and calcium signaling pathways were recently implicated in both CD and hypertension, suggesting a possible relationship between these two complex diseases (Shahbaba, B. *et al.*, 2012).

To illustrate further how biological pathways are involved in disease pathogenesis, Wang et *al.* (2010) manually compiled a pathway centred on the interleukins (IL)−12 and IL−23, which is important pathway for CD (Wang, K. *et al.*, 2010). Figure 1.10 only shows the main proteins in this pathway, which have been previously identified as being associated with Crohn's disease

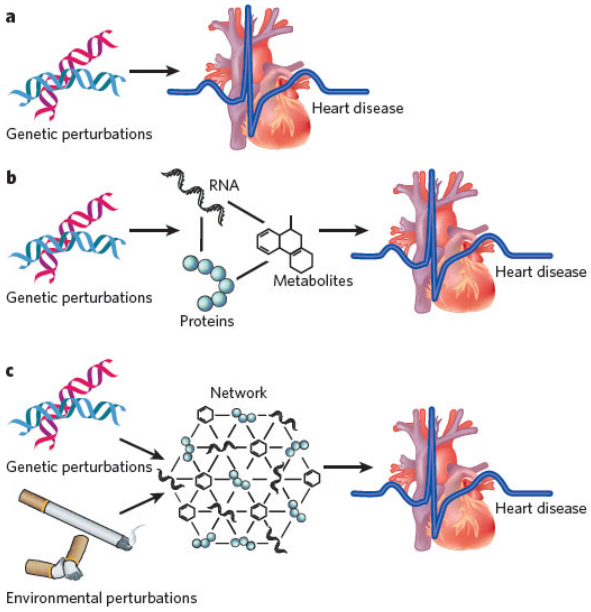

Figure 1.9: Causal relationship models.
(a) Classical genetic association view that identified variations in DNA correlate with disease state or with quantitative traits associated with disease.
(b) DNA variations do not lead to disease on their own but, instead, lead to changes in molecular traits that go on to influence disease risk. By layering in molecular phenotypes as intermediate phenotypes, causal relationships between genes and disease can be established directly.
(c) Disease gene networks sense constellations of genetic and environmental perturbations. Therefore, a more realistic model is one in which constellations of genetic and environmental perturbations affect molecular states of networks that in turn affect disease risk. From (Schadt, E. E., 2009).

in published GWA studies. Only three genes at two loci (IL12B, IL23R and IL12RB2) showed genome-wide significant associations in the GWA study (Barrett, J. C. *et al.*, 2008), however, three genes with moderate association signals in this pathway (JAK2, CCR6 and STAT3) were confirmed as susceptibility genes for CD in replication studies (Barrett, J. C. *et al.*, 2008), and six other genes with modest association signals (STAT4, IL18, IL18RAP, TYK2, IL27 and IL10) were confirmed as CD susceptibility genes in other association studies (Wang, K. *et al.*, 2010).

These examples clearly demonstrate that some genes that may not reach genome-wide significance in any GWA study due to lack of power, may still work together with multiple related genes in the same functional pathway to confer disease. Thus, the susceptibility loci for complex traits are not likely to be randomly distributed (Holmans, P. *et al.*, 2009), and because the underlying genes involved in a disease pathogenesis are likely to be present in the same functional pathways (Sharma, A. *et al.*, 2013). PA approaches may therefore complement the single-marker-based approach for GWAS and provide additional knowledge for disease etiology, by identifying additional susceptibility genes, which could not be detected using conventional methods, using more mechanistic approaches, without any need to reduce the number of possibilities of each GWA study locus to a single gene (Wang, K. *et al.*, 2010).

1.5.2 Methods for pathway-based analysis of GWA studies

Over recent years, several different approaches have been proposed to summarize the significance of a biological pathway from a collection of SNPs from a GWA study, and to adjust for

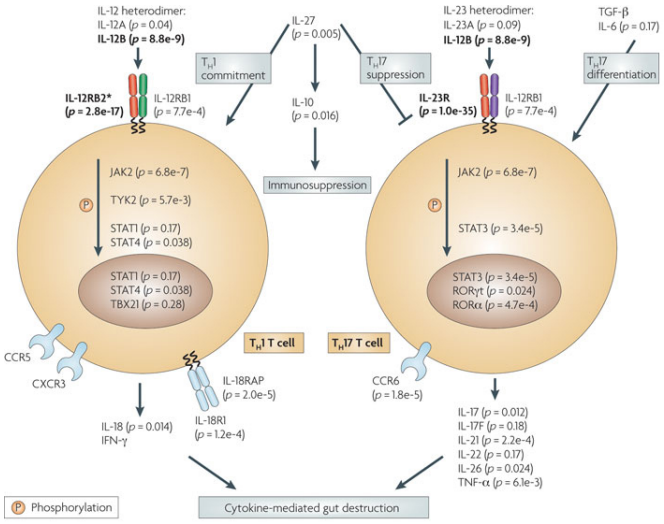

Figure 1.10: Linking biological pathways to complex diseases. For many years, the pro-inflammatory cytokine IL−12, which is mediated by T cells that produced T helper 1 cytokines (T_H1 cells), was believed to be a major player in CD etiology. Recently, studies demonstrated a more important role of the pro-inflammatory cytokine IL−23, which activates a subset of T cells responsible for the production of IL−17 (T_H17 cells), required in the cytokine-mediated gut destruction process. Each gene is annotated with the most significant p-value among all its closest SNPs, and genes with genome-wide significant signals are shown in bold font. From (Wang, K. *et al.*, 2010)

multiple testing at the pathway level (Chen, X. *et al.*, 2010; Guo, Y. F. *et al.*, 2009; Wang, K. *et al.*, 2007; Jia, P. *et al.*, 2011; Jensen, M. K. *et al.*, 2011; Fehringer, G. *et al.*, 2012; Baranzini, S. E. *et al.*, 2009; International Multiple Sclerosis Genetics Consortium, 2013; Subramanian, A. *et al.*, 2005; Tintle, N. L. *et al.*, 2009; Ritchie, M. D. *et al.*, 2001; Yu, K. *et al.*, 2009; Krauthammer, M. *et al.*, 2004; Zhang, K. *et al.*, 2010), and some of these algorithms are publicly available as software packages or web servers (Table 1.8). These PA approaches can be broadly classified into 3 categories with respect to the overall strategy used, including a data mining approach, gene set enrichment analysis approach, and network-based approach (Kraft, P. *et al.*, 2009). In *data mining approaches*, a group of functionally related genes is identified, and multivariable analysis techniques are applied to markers in these genes (Ritchie, M. D. *et al.*, 2001; Yu, K. *et al.*, 2009). More flexible analyses incorporating nonlinear relations between markers and the trait under study may be applied to uncover additional more plausible evidence of association. The major challenge of these approaches emerges when searching over data sets with a large number of markers, which can ultimately lead to substantial downwardly biased p-values and overestimates of the precision of predicted trait values based on the fit model in a new data set (Kraft, P. *et al.*, 2009). In *Gene set enrichment analysis*, a list of markers ranked with respect to their p-values is tested for enrichment of genes in particular functional pathways (Guo, Y. F. *et al.*, 2009; Wang, K. *et al.*, 2007; Zhang, K. *et al.*, 2010). The major challenges for these approaches reside in mapping markers to genes in a many-to-many scenario. Whether a marker is assigned to a single or several genes, and how it is assigned may greatly affect the results. In addition, these approaches tend to favor pathways with several large genes to the detriment of pathways containing only small genes (Kraft, P. *et al.*, 2009). *Network-based approaches* generally consider the set of implicated genes from GWA studies, and attempt to identify nonrandom functional groupings among them (Jia, P. *et al.*, 2011; Raychaudhuri, S. *et al.*, 2009). Though these approaches present a promising aspect as no prior specification of the

set of functional pathways to be analyzed is required, they are also subject to several pitfalls; the biological networks used are not always complete, and sometimes inconsistent, limited by the scope of human knowledge; and the physical clustering of genes with similar functions can create spurious observations that several genes in a pathway are associated with the trait under study (Kraft, P. *et al.*, 2009).

Table 1.8: Publicly available packages for PA on GWAS data sets.

Name	Strategy	Input data	Analysis strategy	URL/REF
MDR	Data mining	Raw genotype	Perform multifactor dimensionality reduction (MDR) to detect potential PPI in case-control studies.	www.cran.r-project.org/web/packages/MDR/ (Ritchie, M. D. *et al.*, 2001)
ARTP	Data mining	Raw genotype	Compute gene and pathway p-values using the Adaptive Rank Truncated test.	www.dceg.cancer.gov/tools/analysis/artp (Yu, K. *et al.*, 2009)
GenGen	GSEA	Raw genotype	Assign the best p-value of SNPs to a gene, then compute Kolmogorov-Smirnov-like enrichment score for a pathway.	www.openbioinformatics.org/gengen/ (Wang, K. *et al.*, 2007)
mSUMSTAT	GSEA	SNP p-values	Similar to GenGen but the pathway is calculated by averaging χ^2 test statistics within each pathway.	www.jurgott.org/linkage/sumstat.html (Fehringer, G. *et al.*, 2012)
GRAIL	Network-based	Raw genotype	Build networks using published abstracts, then identify putative relationships among genes that do not have a single cocitation	www.broad.mit.edu/mpg/grail/ (Raychaudhuri, S. *et al.*, 2009)
dmGWAS	Network-based	SNP p-values	Identify significant PPI modules and candidate genes using a dense module searching method on the GWAS data set and PPI network.	bioinfo.mc.vanderbilt.edu/dmGWAS.html (Jia, P. *et al.*, 2011)

In addition to the principal strategy used, these PA approaches can also be classified into two types with respect to whether the required input data sets are individual SNP genotypes or simply a list of SNP p-values from a GWA study (Wang, K. *et al.*, 2010). The first type, the *p-value enrichment approach*, determines whether a given group of p-values of SNPs or genes is enriched in association signals (Jia, P. *et al.*, 2011; Fehringer, G. *et al.*, 2012). Many of these approaches typically use a p-value cutoff of $p < 0.05$ or $p < 0.001$, meaning, the results are partly dependent on the user-specified cutoff (Figure 1.11a). The second approach, the *raw genotype approach*, aims to derive test statistics at gene and pathway levels using individual SNP genotype data (Ritchie, M. D. *et al.*, 2001; Zhang, K. *et al.*, 2010). This approach requires raw genotypes for phenotype permutation to adjust for the test statistics at both gene and pathway levels, using for instance a two-step correction procedure (Wang, K. *et al.*, 2007).

Though *p-value enrichment approaches* are straightforward in eliminating many practical challenges related to data analysis and data sharing, in addition to the fact that raw genotypes may not always be easily available and permutation procedures computationally expensive, one advantage of *raw genotype approaches* is the ability to maintain the correct LD pattern among neighboring SNPs while permuting (Figure 1.11b).

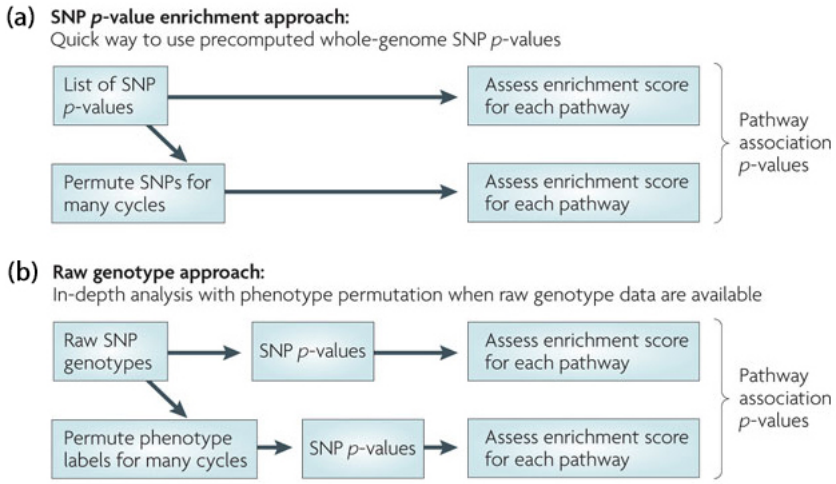

Figure 1.11: Types of PA approaches. A summary of the two main PA approaches for GWA studies based on the input data. From (Wang, K. *et al.*, 2010)

1.5.3 Testing the null hypothesis: competitive and self-contained methods

Another criteria for categorizing PA approaches of GWAS data is based on the way the null hypothesis can be tested, that is, PA approaches can also be divided into competitive (or enrichment) and self-contained (or association) methods. The differentiation between these two methods is important, and key differences between these two approaches emanate from the null hypothesis stated (Fridley, B. L. *et al.*, 2011).

(a) Competitive methods: These methods typically identify first SNPs (or genes) that are significantly associated with the trait studied, and then evaluate whether the significantly associated SNPs are likely to cluster in predefined pathways. Thus, they compare the frequency of significantly associated SNPs in a given pathway P with the frequency of significantly associated SNPs among all genes outside the pathway P (Figure 1.12b). More specifically, they aim to verify the null hypothesis:

H_0: SNPs/genes in a given pathway are associated with the phenotype as much as SNPs/genes outside the pathway.

A commonly used competitive method is the GSEA that uses a modified Kolmogorov-

Smirnov statistic to test the null hypothesis (Wang, K. *et al.*, 2007; Subramanian, A. *et al.*, 2005). Another commonly used approach for competitive PA uses the Fisher's exact test to compare the proportion of associations surpassing some predefined significance level within the pathway P, to the proportion of such association signals outside the pathway P (Holmans, P. *et al.*, 2009). Some examples of competitive PA methods include GenGen (Wang, K. *et al.*, 2007) and i-GSEA4GWAS (Zhang, K. *et al.*, 2010) One major challenge of the Fisher's exact test, and similar methods, is the dichotomization of SNP association signals into significant and non-significant based on a predefined significance level, ignoring information regarding the strength of the association (Fridley, B. L. *et al.*, 2011). Another method based on GSEA, with several improvements of the Kolmogorov-Smirnov statistic, is the Get Set Analysis (GSA) algorithm, which uses a *maxmean statistic* M for the enrichment score, potentially leading to greater power compared to the Kolgomorov-Smirnov test (Hoh, J. *et al.*, 2001; Wu, M. C. *et al.*, 2009), given by:

$$M = max\left\{\left|\frac{\sum_{i=1}^{m} I\left(t_i > 0\right) t_i}{m}\right|, \left|\frac{\sum_{i=1}^{m} I\left(t_i < 0\right) t_i}{m}\right|\right\}, \tag{1.27}$$

where m is the number of genes in the pathway P and t_i is the test statistic of the i-th gene in P. The significance of the M statistic is evaluated using the permutation test involving both genes and class labels.

(b) Self-contained methods: In contrast to competitive methods, self-contained or association approaches only analyze association of SNPs/genes in the pathway P of interest with the phenotype, without taking into account the remaining SNPs/genes. These methods test the null hypothesis versus the alternative hypothesis stated respectively as follows:

H_0: No SNPs/genes in the pathway P are associated with the phenotype.

H_a: SNPs/genes in the pathway P are associated with the phenotype.

A commonly used self-contained method is the Global test, which uses generalized linear models to express the relationship between the expression of genes in P and the phenotype, that is, if the pathway P can be used to predict the clinical outcome, its expression pattern must differ for different outcomes (Wu, M. C. *et al.*, 2009). Let Y be the outcome (phenotype) of interest, continuous or 1/0 for case/control status, and let X be the $n \times m$ matrix of genes association signals for the pathway (with n the number of samples), with x_{ij} the gene association signal of the j-th gene of the i-th sample, the global test is induced by a regression model ("Competitive and self-contained gene set analysis methods applied for class prediction"), such that

$$g(E(Y_i|\beta)) = \alpha + \sum_{j=1}^{m} x_{ij}\beta_j, \tag{1.28}$$

where g is the logit function for case/control status, and α is an intercept, with sample class label Y_i, and β_j is the coefficient for gene j. Then, testing the H_0 that no genes in the pathway P are associated with the phenotype is equivalent to testing:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0. \tag{1.29}$$

Because the sample size is in most cases relatively small compared to the pathway P size, an additional assumption that the coefficient $\beta_1 \ldots \beta_m$ are *iid* (independent, identically distributed) with mean 0 and variance τ^2 is made, which simplifies the hypothesis and make the global test feasible (Wu, M. C. *et al.*, 2009).

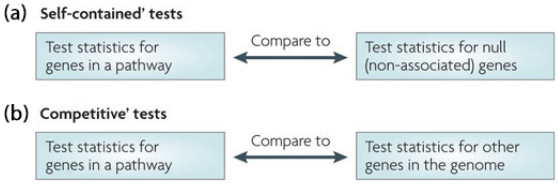

Figure 1.12: Null hypothesis testing in PA. Competitive and self-contained methods for testing the null hypothesis in PA for GWA studies. From (Wang, K. *et al.*, 2010).

Because of differences between these two approaches, the appropriate approach should be selected based on thorough considerations of the null and alternative hypotheses being tested, as well as the constraint imposed by available data. Thus competitive methods can not be applied in studies that consider only a few candidate pathways, whereas self-contained methods can be applied to candidate gene association studies, and in a GWA study with accounted genomic inflation (Fridley, B. L. *et al.*, 2011).

1.5.4 Accounting for gene-level association: one-step and two-step methods

When it comes to whether or not gene-level evidence of association is taken into account when combining the evidence of association in a pathway, two approaches, the *one-step* approach and the *two-step* approach, can be considered (Fridley, B. L. *et al.*, 2011). In the *one-step* approach, all the SNPs in a pathway are used without any consideration of gene-level effects. The *two-step* approach first uses SNPs in each gene to assess the association with the gene, and then combines gene-level tests to evaluate the association of the phenotype with the pathway (Hoh, J. *et al.*, 2001). Different methods are generally used to combine association signals of SNPs to assess the association of the gene with the phenotype, these include using the minimum SNP-specific p-value as the p-value for a gene, using a summary measure of all SNP-specific p-values within the gene, or modeling the simultaneous effects of all SNPs in the gene on the phenotype (Fridley, B. L. *et al.*, 2011). Many PA methods for GWA studies have implemented a two-step approach taking the minimum p-value (or the maximum test statistic) observed for all the SNPs in a gene (Wang, K. *et al.*, 2007; Yu, K. *et al.*, 2009). However, when several distinct SNPs in

the gene contribute to the overall association signal, and all have modest effect on the phenotype, using the minimum p-value may not be the best or most powerful approach to capture such information. In addition, larger genes with more SNPs are likely to have smaller minimum p-values compared to smaller genes with fewer SNPs (Wang, K. *et al.*, 2007; Jia, P. *et al.*, 2012; Fridley, B. L. *et al.*, 2011). Nevertheless, numerous studies have assessed the performance of various methods for combining association signals at the gene-level, which may provide guidance for the completion of a two-step correction in PA for GWA studies (Lehne, B. *et al.*, 2011).

Both the one-step and two-step methods have advantages and disadvantages, and their efficiency depends mostly on the underlying disease-causing mechanisms, which are unfortunately unknown, as well as on the genetic models, for which some study results have indicated that in general, the two-step approach may be more powerful than the one-step approach when the self-contained hypothesis is assessed, and the LD between SNPs needs to be accounted for (Fridley, B. L. *et al.*, 2011).

1.5.5 Impact of LD and adjustment of association significance for pathway size

Because of LD among SNPs in genes composing a pathway, independence between the SNPs can not be assumed in the evaluation of the significance of the pathway, though much less non-negligible correlation may be expected among gene-level p-values. Most PA approaches use permutation procedures to estimate summary adjusted p-values for given pathways. However, for p-value enrichment-based approaches, permutation of SNPs typically disrupts LD patterns between SNPs, thus not generating the correct null distribution. Yet, even in the case of the raw genotype-based approach, permutation of phenotypes (binary or quantitative traits) still introduces some biases, because, when performed, the 'background' distribution reflects the case where none of the SNPs or genes is associated with the phenotype, whereas in practice, for any phenotype, a proportion of SNPs will be genuinely associated with the phenotype in unpermuted data sets (Wang, K. *et al.*, 2010).

As discussed previously in Subsection 1.5.4, ignoring gene size when assessing gene association signal, as well as when testing for pathway association, can lead to inflated type 1 error rates. Permutation procedures can also be used to adjust for size of the pathway and potential bias introduced by PA methods such as the minimum p-value or maximum test statistic for evaluating gene association signal from SNPs in the gene, carefully considering the null and alternative hypothesis and PA methods (competitive or self-contained) being applied. Two-step methods can involve jointly modeling the effects of SNPs within a gene followed by the combination of gene-level association signals to evaluate the association of the pathway with the phenotype. Despite the fact that much less correlation may be expected between the gene-level p-values, a

much lower but non-negligible correlation between genes in a pathway may exist (Fridley, B. L. *et al.*, 2011). Therefore, permutation methods are still crucial in pathway-based approaches for GWA studies, in particular to assess any significant results.

1.6 Challenges and considerations

Pathway-based approaches for GWA studies are becoming an invaluable tool to enable more powerful association tests, and to help enlighten our understanding of disease susceptibility, however a number of pitfalls and challenges, some discussed in this review, limit their practical use. While methods and tools for pathway analysis are dynamically growing in number, several sources of bias must be carefully considered and addressed, including the capacity for strongly associated markers to drive pathway association, the possible effects of SNPs being assigned to multiple genes, the 'history' of the accumulation of mutations that can influence the likelihood of individual genes to show association with the phenotype, and more specifically bias introduced by gene size as well as the bias from pathway size, less commonly addressed in many current methods (Wang, K. *et al.*, 2010).

Numerous large-scale strategies have been developed to study complex diseases, integrating large-scale data derived from diverse strategies for identifying structure and function of genes, including genomics, transcriptomics, proteomics and metabolomics. The integration of such information into pathway-based analysis would provide a fertile process for exploring more sensitive and powerful analysis of GWAS data sets (Ramanan, V. K. *et al.*, 2012). Network approaches, which incorporate biological network structures such as hubs and motifs into the analysis of GWAS data sets, will be vital for this integration in the years to come.

CHAPTER 2

ancGWAS: AN IMPROVED METHOD FOR GENE-GENE INTERACTION ANALYSIS OF GENOME-WIDE ASSOCIATION STUDY DATA FOR HOMOGENEOUS AND ADMIXED POPULATIONS

Genes can influence each other at the level of enhancement or hindrance. The effect can occur directly at the genomic level where a gene could code for a protein that prevents transcription of another gene, alternatively, the effect can be at the phenotypic level, where a pair of genes can interact to produce a specific phenotype. Thus, interactions can play critical roles in the cause of disease. In addition, for a given disease, risk genes may differ in different individuals, but are more likely to lie in the same pathway. Identifying pathways by incorporating prior biological information on multiple genes associated with a disease may allow us to more easily unravel the pathogenesis of complex diseases. Therefore, as discussed in previous chapters, the single-marker-based approach of current GWA studies alone may not be sufficient (see Section 1.3.2). Meta analysis, which aims to pool information from multiple GWA studies to increase the chances of finding associations with small effect sizes, has been successfully used and has increasingly helped in identifying additional susceptibility loci (Han, B. *et al.*, 2011; Cantor, R. M. *et al.*, 2010). A number of additional alternative approaches have been proposed as a complement to the single-marker-based approaches (See Section 1.5). Examining the combined effects of genes by detecting genetic signals beyond single gene polymorphisms may increase our ability to fully characterize the susceptible genes and the genetic structure of complex diseases (See Section 1.5.1). A suggested new paradigm for GWA studies (Jia, P. *et al.*, 2011; Cantor, R. M. *et al.*, 2010) involves incorporating both the association signal from a GWA study and human PPI information to test the combined effects of SNPs and search for significantly enriched sub-networks for a particular complex disease. This approach is based on combining $p-$values from standard GWA studies for correlated SNPs into an overall significance level for

a gene, and combining $p-$values for the genes in a pathway into an overall significance level to investigate the association of a pathway with the disease (Jia, P. *et al.*, 2011). However, in many cases, SNPs within genes and genes within pathways are correlated, and most methods do not account for this dependency, but rather assume genes or SNPs to be independent and uniformly distributed under a null hypothesis. The violation of dependent assumptions in these methods may generate erroneous results. Here, we present an algebraic graph-based method (ancGWAS) to identify the most significant sub-networks that may explain ethnic differences in complex disease risk, both in recently admixed or non-admixed populations by integrating the association signal from GWA study data, the local ancestry (if an admixed population) and polymorphism LD into the human PPI network. The ancGWAS method accounts for the correlation that exists between SNPs within or between genes and between genes within pathways and tests for signals of unusual difference in excess/deficiency of ancestry in admixed populations. In addition, this method introduces flexibility in estimating gene and sub-network-specific ancestry.

2.1 Methods and implementation

2.1.1 Combining $p-$values at the gene level

One challenge of pathway-based analysis methods for GWA study data sets is how to represent gene association with the phenotype (Peng, G. *et al.*, 2010; Casci, T., 2007). In the gene expression field from which the idea of pathway-based analysis of GWA study data sets is borrowed, gene expression arrays yield a single value of the expression level for each gene, whereas in GWA studies each gene can be represented by multiple SNPs, some of which are correlated (Shahbaba, B. *et al.*, 2012). Some PA approaches use the minimum SNP-level signal (maximum statistic) within a gene to represent the gene (Guo, Y. F. *et al.*, 2009; Wang, K. *et al.*, 2007; Jensen, M. K. *et al.*, 2011), however, when multiple distinct SNPs within a gene contribute to the overall association signal, this method can fail to detect additive effects among SNPs with moderate individual association, therefore may no longer be the best statistic to capture such information. This approach also introduces biases, as larger genes are more likely to contain more-significant statistics, and enrichment statistics for pathways containing large genes will be inflated (Ramanan, V. K. *et al.*, 2012; Wang, K. *et al.*, 2007). Therefore, it is not yet clear what the best strategy is to combine statistics for multiple SNPs within a gene into a single value for the gene. In ancGWAS, we implemented three additional methods including Fisher's test, Sime's test, and the Smallest gene-wise FDR, for combining statistics from multiple correlated SNPs, by estimating an overall significance level or summary statistics to represent a gene.

Let us assume multiple SNPs within a gene, all contributing to the overall association of the gene. We assume an independent and uniform distribution of $p-$values p_i for the corresponding

test statistic T_i, testing the $i-$th marker M, under the null hypothesis, although this assumption of independence may be violated because of LD among SNPs within the gene. If we consider a continuous monotonic function H, then a transformation of the p-value is defined as

$$Z_i = H^{-1}(1 - p_i)$$

To combine $p-$values of all SNPs within a gene, we implemented four methods including the maximum test statistic (Sidak's combination test), the Fisher's combination test, the Sime's combination test and the FDR method (Peng, G. *et al.*, 2010).

(a) **Sidak's combination test:** Let us consider only the SNP with the maximum test statistic (the best SNP), we can define the statistic $Z_B = p_{(1)}$, which is distributed as $P(Z_B \leq w) = 1 - (1 - w)^K$, with K the number of independent SNPs.

(b) **Fisher's combination test:** The statistic to combine K independent p-values or to combine information from K SNPs is given by

$$Z_F = -2 \sum_{i=1}^{K} \log p_i,$$

which follows a χ^2_{2K} distribution.

(c) **Sime's combination test:** Let p_i be ordered as $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)}$. The combined $p-$value is given by

$$P_s = \min_i \left\{ \frac{k p_{(i)}}{i} \right\}.$$

(d) **False discovery rate method:** Let $\mathbf{F}(\alpha)$ be the expected proportion of tests yielding a $p-$value less than or equal to α. Suppose a set of $\mathbf{pd} = \{p_1, \ldots p_k\}$ with d distinct $p-$values, with $\tilde{p}_1 < \tilde{p}_2 < \ldots < \tilde{p}_d$. Let m_j be the number of $p-$values equal to $\tilde{p}_j$ among the set of $p-$values $\mathbf{p}$. Then the estimate of the expected proportion $\mathbf{F}(\alpha)$ is given by

$$\tilde{\mathbf{F}}(\alpha) = \frac{1}{k} \sum_{j=1}^{d} \mathbf{I}(\tilde{p}_j \leq \alpha) m_j,$$

where $\mathbf{I}$ is an indicator function. For a one-sided test (χ^2-test, or trend test) consider $\pi = \min(1, 2\tilde{a})$, and for a two-sided test consider $\pi = \min(1, 2\tilde{p})$, where $\tilde{p} = \frac{1}{k} \sum_{i=1}^{k} p_i$, $\tilde{a} = \frac{1}{k} \sum_{i=1}^{k} a_i$, and $a_i = 2 \min(p_i, 1 - p_i)$. The estimate $\hat{v}(\alpha)$ of the expected proportion $v(\alpha)$ of tests resulting in a false positive when α is used as the $p-$value threshold to determine significance is expressed as

$$\hat{v}(\alpha) = \hat{\pi}\alpha,$$

where $\hat{\pi}$ is the estimate of the proportion π of tests with the true hypothesis. Then, the FDRs are expressed as ratios of the form

$$t_{(i)} = \frac{\hat{v}\left(p_{(i)}\right)}{\hat{\mathbf{F}}\left(p_{(i)}\right)},$$

and

$$q(i) = \min_{j \geq i} \{t_j\},$$

where $q_{(1)} \leq q_{(2)} \leq \ldots \leq q_{(m)}$ are the ordered false discovery rates. Finally, $q_{(1)} = \min\left\{t_{(j)}\right\}$ is the false discovery rate for the gene.

2.1.2 Constructing the LD-weighted PPI network

SNPs, their associated local ancestry, ancestral population minor allele frequencies and GWA study $p-$values are assigned to a given gene if they are located within the gene's primary transcript or 20 kilobases (kb) downstream or upstream. Note that, this distance can still be decided or adjusted in ancGWAS according to the genetic architecture of the population under study, given that some genetic variants that may influence the disease might be located much further than 20 kb downstream or upstream of the gene. If the most significant SNP from GWAS is located out of this boundary, then all the SNPs within this distance will be included in the analysis. If a SNP is assigned to multiple genes due to overlapping flanking regions, the closer gene is chosen according to a specified boundary cut-off. In a similar manner, a given sub-network (discussed in the next section) can include SNPs associated with each of its genes.

Gene-gene dependency is complex and is crucial for the understanding of different biological mechanisms (Liang, H. *et al.*, 2007; Wu, J. *et al.*, 2009; Peng, G. *et al.*, 2010). Such dependencies between genes are attributed to physical interactions between encoded proteins or between a protein and a gene, or as a consequence of coregulation of some transcription factors (Gao, X. *et al.*, 2009). For instance, in human cancer, the ataxia-telangiectasia mutated (ATM) kinase, a major regulator of the cellular response to DNA double-strand breaks, is required for p53 tumor suppressor accumulation and phosphorylation in response to increased Myc activity, by promoting apoptosis (Pusapati, R. V. *et al.*, 2006). Thus if we can identify all gene dependency pairs underlying a specific phenotypic change, we can better understand the biological mechanism related to the state change of the phenotype. From the genotype data, we weight the known pairwise protein-protein interactions (PPI) network (Wu, J. *et al.*, 2009; Cowley,

M. J. *et al.*, 2012) with a weighted correlation. This is accomplished by computing an overall LD from pairwise LD between SNPs of each given pair of interacting genes. In this manner, the PPI network is weighted with a correlation estimated from the population genotype data. This may be useful in efficiently breaking the PPI network into different sub-networks (see the next section) and helpful in combining $p-$value approaches that account for dependency of neighbouring genomic markers within/between genes.

Given two different sets of SNPs $S^a = \{s_i\}_{i=1,2,\ldots,m}$ and $S^b = \{s_j\}_{j=1,2,\ldots,n}$, $(s_i \neq s_j,\ i = 1,2,\ldots,m$ and $j = 1,2,\ldots,n)$ associated with genes g_a and g_b, the pair-wise LD between SNPs in S^a and S^b are computed using the r^2 measure (See Subsection 1.1.3) and a combined LD can be obtained. Here, we provide three approaches for weighting these interactions (i.e. estimating the combined LD at the gene level).

(a) **Case 1. ClosestLD**: Considering each SNP $s_i, (i = 1,2,\ldots,I)$ along the genome is assigned to its closest gene g_k, $k = (1,2,\ldots,K)$, thus each pair-wise LD, $LD_{s_i s_j}$, $s_i \neq s_j$ $(i,j \in 1,2,\ldots,I)$ is considered as a pseudo correlation $r_{g_l g_k}$ of genes g_a and g_b, for $a,b \in 1,2,\ldots,K$.

$$r_{g_a g_b} = LD_{s_i s_j}. \tag{2.1}$$

(b) **Case 2. ZscoreLD**: We assume sets of SNPs $S^a = \{s_i\}$ and $S^b = \{s_j\}$, $s_i \neq s_j$, $(i = 1,2,\ldots,m;\ j = 1,2,\ldots,n)$ are assigned to genes g_a and g_b; and the pairwise LD of SNPs between g_a and g_b are independent. Because the distribution of the LD is not normal, from $s_i \neq s_j$, we compute the average z-transforms of LD from all possible combinations of pairs of SNPs between genes g_a and g_b. The z-transforms of LD are normally distributed with mean 0 and variance 1. We compute the combined LD between two genes g_a and g_b (Choi, S. C., 1977) as follows

$$r_{g_a g_b} = tanh\left(\frac{\sum_{i\neq j}^{N} tanh^{-1}\left(LD_{s_i s_j}\right)}{N}\right). \tag{2.2}$$

(c) **Case 3. maxLD**: Alternatively to the case above, if SNPs between a given pair of genes are dependent or correlated, we consider the maximum among all possible N pairs of SNPs between the pair of genes given by

$$r_{g_a g_b} = \max_{i\neq j}(tanh^{-1}\left(LD_{s_i s_j}\right)). \tag{2.3}$$

The combined LD is used as the weight of the edge between g_a and g_b genes in the PPI network.

2.1.3 Searching for sub-networks using centrality measures

Genes interact in large networks in all living organisms, and some genes in the network are more important or central than others. Highly connected genes in PPI networks can be functionally important and the removal of such nodes is related to lethality (Liang, H. *et al.*, 2007). We consider our weighted PPI network as an undirected network, $G = (V, E)$, where V is the set of n genes as nodes and E is the set of edges as interactions found between genes weighted using gene-correlation (See Subsection 1.4.1). To cluster G into sub-networks, we analyse the general topological properties of G and quantify the usefulness of each gene in G using their centrality scores; closeness, betweenness, degree or eigenvector. Let us first define the following centrality measures.

(a) **Degree centrality:** The degree centrality C_d of a gene g in an undirected graph G is the number of genes in the network interacting with it. In terms of the adjacency matrix $\mathcal{A} = (a_{uv})_{1 \leq u,v \leq n}$, which is symmetric for an undirected graph, the degree centrality of a node $u \in V$ (u as a column or a row in the adjacency matrix) is simply the sum of components in the column or row corresponding to the node u, i.e. $\sum_{v \in V} a_{uv} = C_d = \sum_{v \in V} a_{vu}$ (Mazandu, G. K. *et al.*, 2011). The degree centrality provides an indicator of the influence of a gene in a biological system and indicates whether the gene plays a key role in the functioning of the system. C_d is also used, for instance, to correlate the degree of a gene in the network with the lethality of its removal. It could also indicate a central role in gene expression (transcription factors), diversification and turnover (small GTPases), or signaling module assembly (docking proteins) (Scardoni, G. *et al.*, 2009).

(b) **Closeness centrality measure:** The closeness centrality of a node u, $C_c(u)$ in a network G, is the inverse of the average distance to all other nodes connected to it (Mazandu, G. K. *et al.*, 2011), and is expressed as

$$C_c(u) = \frac{1}{\frac{1}{m-1} \sum_{v \in V_u} dist(u, v)}, \tag{2.4}$$

where V_u is the set of nodes connected to u, $m = |V_u|$ the number of nodes in V_u and $dist(u, v)$ is the shortest communication between u and v in the network. The closeness centrality measure provides an indication on how a gene is functionally relevant for other genes in terms of accessing information via or propagating information to other genes in the connected part of the network containing that gene. Thus, a gene with high closeness, compared to the closeness of the whole network, may be central to the regulation of other genes (Bjorn, H. J. *et al.*, 2008).

(c) **Shortest path betweenness centrality measure:** Let us denote σ_{uv}, the number of shortest paths between genes u and v, and $\sigma_{uv}(t)$ the number of shortest paths between

u and v in the network G using t as an interior node, for t, u, $v \in V(G)$. The rate of communication between u and v, δ_{uv} that can be controlled by an interior gene t, is given by

$$\delta_{uv}(t) = \frac{\sigma_{uv}(t)}{\sigma_{uv}}, \tag{2.5}$$

if $\sigma_{uv} = 0$, then we set $\delta_{uv} = 0$. The shortest path betweenness centrality $C_{spb}(t)$ is given by

$$C_{spb}(t) = \sum_{u \in V \wedge u \neq t} \sum_{v \in V \wedge v \neq t} \delta_{uv}(t). \tag{2.6}$$

In protein networks, the shortest path betweenness centrality of a protein may determine its relevance in holding together communicating genes and the capability of a protein to enable communication between distant genes (Bjorn, H. J. *et al.*, 2008).

(d) **Eigenvector centrality measure:**

The eigenvector centrality measure assesses the usefulness or the weight of functional connections of genes and can only be considered as a measure of centrality if genes are ranked based on their participation in different sub-networks. The eigenvector centrality measure assigns relative weights to all genes in the network based on the fact that connections to high-weighted genes contribute more to the weight of the gene target. This means the weight or the contribution x_u of the gene u to the functioning of the system is proportional to the sum of the scores of all genes v connected to u, expressed as

$$\sum_{(u,v) \in E} x_v = \lambda x_u, \tag{2.7}$$

λ is a constant of proportionality and x_z denotes the contribution of gene z.
In terms of the adjacency matrix $\mathcal{A} = (a_{uv})_{1 \leq u,v \leq n}$, we have

$$\sum_{v=1}^{n} a_{uv} x_v = \lambda x_u, \tag{2.8}$$

with n the number of genes in the network. It turns out that λ is simply an eigenvalue of the adjacency matrix and the vector of contributions of genes is the eigenvector associated with λ (Bjorn, H. J. *et al.*, 2008). In fact, it can only be considered as a measure of centrality if nodes are ranked according to their participation in different network subgraphs; in a regular graph for instance, all the components of the main eigenvalue are identical (Rodriguez, J. A. *et al.*, 2007).

Genes that are central in association with complex disease susceptibility are considered to be centres of biological sub-networks, and are linked to other genes in that sub-network via few steps (paths in the network) (Jia, P. *et al.*, 2011). These centres are structural hubs with

centrality scores beyond a certain threshold value. Biological topological property tests of a biological network can guarantee the aforementioned. Let us denote $o(G)$ the order, $s(G)$ the size of G and SP_{mean}, the shortest-path mean from every node to every destination within the network G. Note that the cut-off of different network centrality measures are estimated using general topological properties of the network (Mazandu, G. K. *et al.*, 2011). For the betweenness measure, the cut-off is the total number of shortest paths expected in the network, which is approximately $o(G) * SP_{mean}$. For the closeness metric, as defined in equation (2.4), the cut-off is $1/SP_{mean}$ and that of the degree centrality measure is the number of expected interacting genes of a gene in the network, which is $s(G)/o(G)$. In the case of the eigenvector measure, the cut-off is the mean value of the weight or contribution vector of all genes in the network. We perform the steps in Algorithm 2.1 to identify sub-networks using centrality scores of each gene.

Algorithm 2.1 ancGWAS Clustering Strategy

(1) Given network G, find structural hubs and connected components;

(2) For each gene, compute the betweenness, the closeness and the eigenvector scores;

(3) For each centrality score, compute the cut-off for central genes of sub-graphs BetOf, ClosOf, DegOf and EigOf;

(4) Consider a gene as a hub if its score is greater than or equal to the corresponding cut-off;

(5) Consider a gene as a central gene if it is a hub for all the four scoring measures in step (3);

(6) For each central gene, search its neighbours given a step n or the mean shortest path. The central gene and its neighbours constitute a sub-network of G.

2.1.4 Combining $p-$values at the subnetwork level

Combined p-value approaches are commonly utilized for meta-analysis (Han, B. *et al.*, 2011; Cantor, R. M. *et al.*, 2010), as are using neighbouring genomic markers in genetic association studies (Zaykin, D. V. *et al.*, 2002; Jia, P. *et al.*, 2011; Peng, G. *et al.*, 2010), and these have a long history (Folks, J., 1984; Hedges, L. *et al.*, 1985; Loughin, T. M., 2004). Under the null hypothesis, the p-values p_i, $(i = 1, \ldots, L)$ for a test-statistic with a continuous null distribution are uniformly distributed in the interval $[0, 1]$. In this framework, a parametric cumulative distribution function F can be chosen and the $p-$values can be transformed into quantiles according to $q_i = F^{-1}(p_i)$, $(i = 1, \ldots, L)$. The combined test statistic $C^P = \sum_{i=1}^{L} q_i$ is a sum of independent and identically distributed random variables q_i, each of which follows the corresponding probability density function for F. To account for the independent assumption of $p-$values and the correlation of $p-$values among neighboring genomic markers, we implemented

both the Stouffer-Liptak (Liptak, T., 1958) and Fisher's Combined probability (Fisher, R., 1958; Hess, A. *et al.*, 2007) methods accounting for spatial correlations among SNPs within a gene or SNPs within a given sub-network.

(a) **Fisher's combined probability test:** Let $\hat{F}$ be the cumulative χ^2 distribution. Let p_i, $(i = 1, 2, \ldots, L)$ be $p-$values of SNPs associated with a gene or a sub-network. We obtain the combined p-value test statistic (Fisher, R., 1958; Hess, A. *et al.*, 2007) $C^P = -2\sum_{i=1}^{L} \log(1 - p_i)$ which follows a χ^2 distribution with $2L$ degrees of freedom due to the additivity property of independent χ^2. The combined p-value is $p^* = \hat{F}(C^P)$, where $\hat{F}$ is the cumulative distribution function for the χ^2 distribution with $2L$ degrees of freedom. Suppose that p_i are dependent, we can approximate the distribution of Fisher's Combined Probability C^P by a scaled χ^2 distribution such that $C^P \approx c \times \chi^2_f$, where c is a scaling factor and f is the degree of freedom. Let,

$$\begin{aligned} E(C^P) &= E\left(c \times \chi^2_f\right) \\ &= c \times f \\ \text{Var}(C^P) &= \text{Var}\left(c \times \chi^2_f\right) \\ &= 2 \times c^2 f \end{aligned}$$

Solving the above equation with respect to c and f, we obtain,

$$\begin{aligned} c &= \frac{\text{Var}(C^P)}{2E[C^P]} \\ &= \frac{4L + 2\sum_{i<j} Cov[-2\log(p_i), -2\log(p_j)]}{4L} \\ f &= \frac{E(C^P)^2}{\text{Var}(C^P)} \\ &= \frac{2(2L)^2}{4L + 2\sum_{i<j} Cov[-2\log(p_i), -2\log(p_j)]} \end{aligned}$$

Since $C^P \approx \chi^2_{2L}$. The combined p-value for C^p is determined by using the approximating distribution ${}^{C^P}/_c \approx \chi^2_f$. To compute terms in Var $\left(C^P\right)$, we apply approximations described in (Kost, J. T. *et al.*, 2002) by fitting a polynomial regression to the true values using a grid approach ranging values for the degrees of freedom $(9 \leq \nu \leq 125)$ and the auto-correlations σ $(-0.98 \leq \rho \leq 0.98)$. We finally compute a q-value score based on the Benjamini-Hochberg false-discovery correction (Benjamini, Y. *et al.*, 1995).

(b) **Stouffer-Liptak test**

Let F be a cumulative standard normal distribution $\mathcal{N}(0,1)$. It follows that $F(x) = \Phi(x)$, where Φ is a cumulative distribution function of the standard normal, and $q_i = \Phi^{-1}(p_i)$. Each q_i, $(i = 1, \ldots, L)$ follows the probability density function of a standard normal and using the additivity property of independent random variables, the Liptak's combined p-value test statistic (Liptak, T., 1958) is

$$C^P = \frac{\sum_{i=1}^{L} q_i}{\sqrt{L}}$$

and the related combined p-value is $p^* = \Phi(C^P)$. To adjust for $p-$value dependency, we assume that p_i $(i = 1, \ldots, L)$ are correlated according to a positive definite and non-degenerate correlation matrix Σ. It follows that the Cholesky factor C exists such that $\Sigma = CC^T$. We transform the correlated quantiles $Q = q_i$ into independent quantiles $\hat{Q}$ as in (Zaykin, D. V. *et al.*, 2002). We then obtain the transformation $\hat{Q} = C^{-1}Q$, the q_i $(i = 1, \ldots, L)$ are now independent and follow a standard normal distribution (Zaykin, D. V. *et al.*, 2002). Finally, we apply the Stouffer-Liptak test on $\hat{Q}$. As for Fisher's Combined Probability test, here we compute a q-value on a null-model from shuffled $p-$values.

2.1.5 Combining local ancestry at the gene and sub-network levels

The aim is to combine the average locus-specific ancestry ϕ_{mk} from k^{th} ancestral populations (estimated from locus-specific ancestry among I admixed individuals) of SNP $m \in \{1, 2, \ldots, L\}$ associated with a given gene (or to a combined set of SNPs associated with each gene within a sub-network). Each average locus-specific ancestry ϕ_{mk} can be considered as a single observation at SNP m, and we make an assumption that each observation ϕ_{mk} can be approximated to be a normal distribution under neutral drift with mean 0 and empirical variance V_{mk} derived from the distribution of locus-specific ancestry among the N admixed individuals. Such an assumption is commonly acceptable in admixture studies because of the large sample size (Henn, B. M. *et al.*, 2012; Baran, Y. *et al.*, 2012). Here we consider the likelihood of these observations. For easier presentation, we describe the method at the gene level.

Let $\{\phi_{1k}, \phi_{2k}, \ldots, \phi_{Lk}\}$ be the average locus-specific ancestry of SNP $m \in \{1, 2, \ldots, L\}$ associated with gene g_j, $(j = 1, 2, \ldots J)$. Let V_{mk} and W_{mk} be the variance and inverse variance (precision) of ϕ_{mk}, and μ_{jk} be the unknown true gene-specific ancestry of g_j from the k^{th} ancestral population. Assuming ϕ_{mk}, $m \in \{1, 2, ..., L\}$ from the k^{th} ancestral population have similar magnitude at g_j, testing whether $\mu \neq 0$, we can derive μ_{jk} from the maximum likelihood of alternative hypotheses $\mathtt{L}_1$ by solving:

$$\frac{\partial \mathtt{L}_1}{\partial \mu_{jk}} = 0$$

$$\frac{\partial}{\partial \mu_{jk}} \left(\prod_{m=1}^{L} \frac{1}{\sqrt{2.\pi V_{mk}}} \exp\left(-\frac{(\phi_{mk} - \mu_{jk})^2}{2V_{mk}} \right) \right) = 0$$

where μ_{jk} is the unknown true gene-specific ancestry at g_j from the k^{th} ancestral population. It follows that the maximum likelihood estimate of μ_{jk} and its variance ν_{jk} is given by

$$\begin{aligned} \hat{\mu}_{jk} &= \frac{\sum_{m=1}^{L} W_{mk}\phi_{mk}}{\sum_{m=1}^{L} W_{mk}} \\ \hat{\nu}_{jk} &= \frac{1}{\sum_{m=1}^{L} W_{mk}} \end{aligned} \tag{2.9}$$

It is necessary to account for the bias of each observed average locus-specific ancestry (obtained from current locus-specific ancestry inference methods) and accuracy of the gene-specific ancestry estimates when combining local ancestry information. To address this, we use an empirical approach that uses the posterior probability ρ_{jk} that the gene-specific ancestry at gene g_j is unbiased. ρ_{jk} is estimated from the data as the prior weight for each ancestral population. Since the posterior probability ρ_{jk} is estimated using all possible ancestral populations of the admixed population, this approach can be thought of as gathering information from all contributing ancestral populations and distributing back in the form of weight at gene-specific ancestry. Let f_{mk} be minor allele frequency of SNP $m \in \{1, 2, \ldots, L\}$ associated with gene g_j, $(j = 1, 2, \ldots J)$ from the k^{th} ancestral population. We can approximate the precision $\frac{1}{\hat{\nu}_{jk}}$ such as $\frac{1}{\hat{\nu}_{jk}} \approx \eta_{jk} = \sum_{m=1}^{L} N.f_{mk}\,(1 - f_{mk})$, where N is the sample size in the k^{th} ancestral population. Now let $\hat{X}_k$ and $\hat{\varepsilon}_k$ be the gene-specific ancestry (see equation 2.9) and the precision from other $K - 1$ ancestral populations $l \neq K \in 1, 2, \ldots K$.

$$\begin{aligned} \hat{X}_k &= \frac{\sum_{l \neq k}^{K} \hat{V}_{jl}\hat{\mu}_{jl}}{\sum_{k \neq}^{K} \hat{\nu}_{jl}}, \\ \hat{\varepsilon}_k &= \frac{1}{\sum_{k \neq l}^{K} \hat{\nu}_{jl}}, \end{aligned}$$

Applying the Baye's theorem, we approximate the posterior probability ρ_{jk}, at each gene g_j as follows,

$$\hat{\rho}_{jk} = \frac{\pi \int_{\mathbb{R}} \mathcal{N}(\phi_{mk}|\theta, \eta_{jk}) p(\theta) d\theta}{(1-\pi)\mathcal{N}(\phi_{mk}|0, \eta_{jl}) + \pi \int_{\mathbb{R}} \mathcal{N}(\phi_{mk}|\theta, \eta_{jl}) p(\theta) d\theta}, \quad l \neq k \in 1, 2, \ldots, K. \quad (2.10)$$

Let $\theta \approx \mathcal{N}(\hat{X}_k, \hat{\varepsilon}_k)$ be an empirical prior on θ. Applying the symmetric propriety of the Gaussian distribution and the fact that the product of Gaussian probability density functions can yield a single Gaussian density function, equation 2.10 becomes,

$$\hat{\rho}_{jk} = \frac{\pi \mathcal{N}(\hat{\mu}_{jk}|\hat{X}_k, \eta_{jl} + \varepsilon_k)}{(1-\pi)\mathcal{N}(\hat{\mu}_{jk}|0, \eta_{jl}) + \pi \mathcal{N}(\hat{\mu}_{jk}|\hat{X}_k, \eta_{jl} + \varepsilon_k)}, \quad l \neq k \in 1, 2, \ldots, K. \quad (2.11)$$

Thus, at gene g_j for $k \neq l \in \{1, 2, \ldots, K\}$, we weight the maximum likelihood estimate of μ_{jk} in equation 2.9 by the posterior probability ρ_{jk} in equation 2.11,

$$\theta_{jk} = \frac{\hat{\mu}_{jk} \sum_{l=1}^{K} \rho_{jl}}{\sum_{l=1}^{K} \rho_{jl}^2}. \quad (2.12)$$

The approach described above is similar to the leave-one-out cross-validation approach usually used in examining statistical prediction (Han, B. *et al.*, 2011).

2.1.6 Testing case-control ancestry difference for gene or sub-network levels

Let Θ_{jk}^{+} and Θ_{jk}^{-}, $(j = 1, \ldots, N)$ be the gene-specific ancestries (for N genes within a given sub-network) estimated from n_1 samples of case and from n_2 samples of control individuals. Assuming $|\Theta_{jk}^{+} - \Theta_{jk}^{-}| \neq 0$, $(j = 1, \ldots, N)$, thus let Γ_j be the rank pairs from smallest absolute difference to largest absolute difference within the sub-network. We use the Wilcoxon signed-rank statistic $\mathcal{W} = |\sum_{j}^{N} sign\left(\Theta_{jk}^{+} - \Theta_{jk}^{-}\right).\Gamma_j|$ (Wilcoxon, F., 1945), a non-parametric test of the null hypothesis that the gene-specific ancestries from cases and controls are the same against an alternative hypothesis. Because N increases, the sampling distribution of $\mathcal{W}$ converges to a normal distribution (Wilcoxon, F., 1945; Siegel, S. *et al.*, 1998), therefore we construct our weighted z-score as follows,

$$Z_{\mathcal{W}} = \frac{(\mathcal{W} - 0.5) \sum_{j}^{N} |\rho_{jk}^{+} - \rho_{jk}^{-}|}{\sigma_{\mathcal{W}} \sqrt{\sum_{j}^{N} |(\rho_{jk}^{+})^2 - (\rho_{jk}^{-})^2|}}, \quad (2.13)$$

where, $\sigma_{\mathcal{W}} = \sqrt{\frac{N(N+1)(2N+1)}{6}}$, ρ_{jk}^{+} and ρ_{jk}^{-} are the posterior probabilities estimated from case and control as defined in equation 2.11 in the section above. The *sign* is an odd mathematical

function that extracts the sign of a real number (Shirokov, Y. M., 1979). Although the p-value can be calculated from enumeration of all possible combinations of $\mathcal{W}$ given N. However, the weighted posterior probability in equation 2.13 is not independent of $\mathcal{W}$, the statistic does not follow a normal distribution. Thus, we compute the p-value using the importance sampling approach as described in (Wasserman, L., 2004; Penny, P. D. *et al.*, 2000) by allowing the sampling distribution to be centered at $\mathcal{Z}_{\mathcal{W}}$.

2.1.7 Characterization of enriched sub-networks

Here we aim at identifying the association between each sub-network (obtained from our network-based clustering approach) S_i, $(i = 1, 2, \ldots, T)$ within $n_1, n_2, \ldots, n_T$ genes and a set P_j, $(j = 1, 2, \ldots, J)$ of human pathways. We obtain $1,047$ annotated pathways from (Feng, Z. *et al.*, 2012) and collected more than 107 annotated pathways from the KEGG (http://www.genome.ad.jp/kegg/), BioCarta (http://www.biocarta.com/), and Ambion GeneAssist Pathway Atlas (http://www.ambion.com/tools/pathway) pathway databases. We downloaded genomic coordinates for all genes from the NCBI ftp-server (www.ncbi.nlm.nih.gov) and retained only entries for the human reference sequence and protein-coding genes. We updated genomic coordinates to the latest assembly using the Lift-Over tool in GALAXY (http://galaxy.psu.edu/). We assign the SNPs located within a gene or less than a particular distance up/downstream distance (for example $< 20kb$) of the gene.

Let α be the intersection between genes within S_i and genes within pathway P_j. Let β be the intersection between genes within S_i and the total genes in P_j, $(j = 1, 2, \ldots, J)$. Let N^* be the intersection between genes in the P_j pathway and the total of genes in P_j, $(j = 1, 2, \ldots, J)$ and M^* be the total of genes in P_j, $(j = 1, 2, \ldots, J)$. We compute the statistic of significance of overlap between sub-network S_i of n_t genes and a given pathway P_j using Z-score (Z_S), which employs the binomial proportions test (Berger, S. I. *et al.*, 2007),

$$Z_S = \frac{\left(\frac{\alpha}{N^*} - \frac{\beta}{M^*}\right)}{\sqrt{\frac{\frac{\beta}{M^*}\left(1 - \frac{\beta}{M^*}\right)}{M^*}}}. \tag{2.14}$$

The approach above does not only score association of overlapping gene sets and a given pathway, it has the advantage of accounting for the network structure of physical interactions between the gene sets through sub-networks.

2.2 Implementation and discussion

2.2.1 Implementation

We implemented the method and algorithm described in Methods (Section 2.1) in ancGWAS. ancGWAS has the advantage of not only using a LD-weighted network, but also has the flexibility to test for signals of unusual excess/deficiency of ancestry, test genes/sub-networks that may explain ethnic differences in complex disease risk, and ancestry proportion at the gene and sub-network levels in admixed populations.

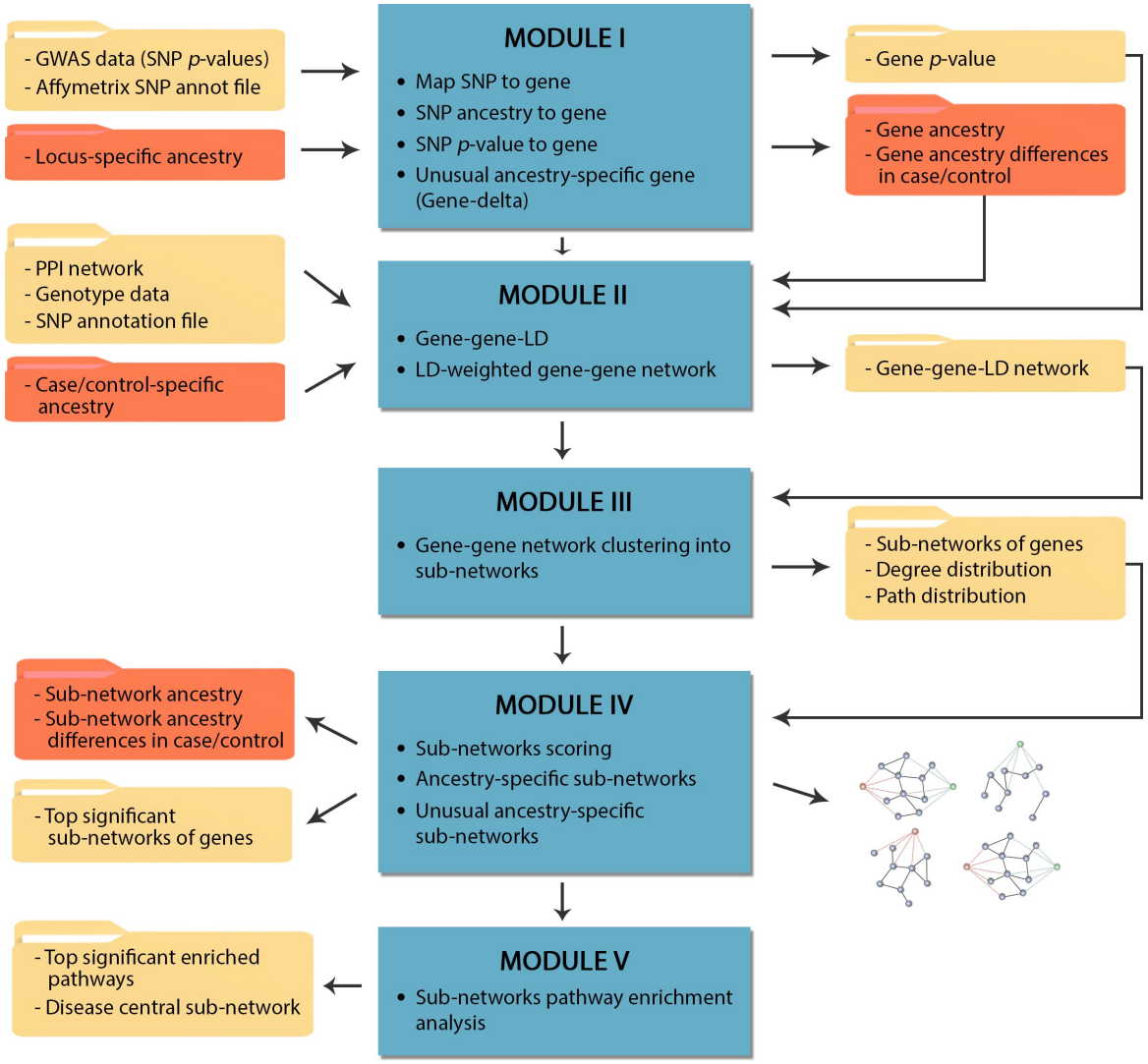

Figure 2.1: ancGWAS workflow. ancGWAS is composed of 5 principal modules (classes), which implement all the methods presented above. They perform, in a step-wise manner, the whole ancGWAS algorithm from reading the input to generating the result, either for a homogeneous or heterogeneous population. Red boxes represent inputs and outputs when the GWA study data set population is heterogeneous, while the yellow boxes are the standard steps for homogeneous, as well as heterogeneous populations.

ancGWAS achieves these through a step-wise strategy, integrating the association signal from GWA study data, the local ancestry and polymorphism pair-wise LD into the human PPI network and examining the topological structure of the resulting network. Written in python using Object Oriented Programming (OOP) abstraction, in a standalone packaging, it also offers an interactive way to process each step of ancGWAS as described in the workflow 2.1. The running time is subject to the size of the generated network, which is a factor of the number of SNPs and mapped genes, and the complexity of the network which, in turn is factor of the number of interactions and other global network properties such as the degree distribution. The running time also increases given that ancGWAS computes the gene-gene correlation from the SNP genotyping data of the GWA study.

2.2.2 Discussion

Pathways are only as good as the annotations on which they are based, therefore, pathway-based approaches for GWA study analysis will continually improve. Here, we introduced ancGWAS, a network-based approach and post GWA study method that integrates the association signal from GWA study data sets, the local ancestry and gene pair-wise LD into the human PPI network, for recently admixed or non-admixed populations. Our method accounts for the correlation that exists between SNPs within a gene and genes within pathways and introduces flexibility in estimating gene-specific and sub-network specific ancestry. It also tests for signals of unusual excess/deficiency of ancestry at the gene and sub-network level, which to our knowledge, is a new contribution in post-GWA study methods. In the next chapter, we assessed ancGWAS through different simulation scenarios, including a simulation of interactive disease loci in an admixed population, and a simulation of a pathway-based case-control association study.

CHAPTER 3

EVALUATION OF ANCGWAS THROUGH SIMULATION OF DISEASE IN NON-ADMIXED AND ADMIXED POPULATIONS

We thoroughly evaluated the performance of ancGWAS using different simulation scenarios, including a simulation of interactive disease loci in an admixed population, and a simulation of pathway-based association study data. We demonstrated through these assessments that ancGWAS outperformed current approaches especially when compared to dmGWAS (Jia, P. *et al.*, 2011), and holds promise for comprehensively examining the interactions between genes underlying the pathogenesis of genetic diseases and also underlying ethnic differences in disease risk. The choice of dmGWAS as a comparison is motivated by the fact that it uses the same strategy for the analysis of GWA study data sets, namely the network-based approach (see Subsection 1.5.2), in integrating GWA study signals into the human PPI network for further analysis. This therefore enables a reasonable comparison between these two approaches, though the final results of ancGWAS could also have been compared to any other existing pathway-based method for analysis of GWA study data sets. In this chapter, we first briefly introduce the dmGWAS method that was used to assess the new proposed method, focusing mostly on the searching strategy implemented in this model, in order to highlight major differences between the two methods. The first assessment based on case-control simulated data of a 4-way admixed population allows us to evaluate how well ancGWAS combines locus-specific ancestries while accounting for the bias of each observed locus-specific ancestry (obtained from current locus-specific ancestry inference methods), and also assess the accuracy of the gene-specific ancestry estimates when combining the observed locus-specific ancestry information. The second assessment allowed us to evaluate the capabilities of ancGWAS to identify important disease genes, based on a case-control data set containing a simulated disease-pathway, and subsequently, compare the performance of ancGWAS with dmGWAS (Jia, P. *et al.*, 2011).

3.1 Materials and methods

3.1.1 Overview of the dmGWAS method

The dmGWAS method uses the dense module searching (DMS) algorithm for scoring modules or sub-networks in massive networks. A module $M \subseteq V$, of a graph $G = (V, E)$, is a set of nodes that have the same neighbors outside the module. Biologically, they define a module as a sub-network within the whole network which has a local maximum proportion of higher test statistics (Jia, P. *et al.*, 2011). The quantitative evaluation of the density of a module with k nodes, is defined by the score

$$Z_m = \frac{\sum z_i}{\sqrt{k}},$$

where z_i is calculated using the p-value of each gene, using the inverse normal distribution function Φ^{-1} given in (Ideker, T. *et al.*, 2002)

$$z_i = \Phi^{-1}(1 - P_i).$$

Thus, if the p-value is small, the score z_i will be large and we seek the module with the higher Z_m score, which would normally be the module with the higher proportion of markers with more significant p-values (Ideker, T. *et al.*, 2002).

To ensure that the score of a module is not too much higher than expected, a normalization is computed using random sets of genes. The same number of genes from the network is randomly chosen $100,000$ times, for a module m of size k. Accordingly, a score $Z_m(\pi)$ will be generated as an estimated background distribution of Z_m, and is given by

$$Z_N = \frac{Z_m - mean(Z_m(\pi))}{SD(Z_m(\pi))}. \tag{3.1}$$

Note that in Equation (3.1), Z_N does not depend anymore on k, meaning that different modules, m_i, can now be compared, independently of their size k_i, therefore the normalized score Z_N of Z_m is used to rank modules.

After scoring modules, a strategy is implemented for searching for dense modules. This procedure is computed depending on two important factors d and r. The parameter d is the constraint distance for which, any node whose the shortest path to another node is greater than this cut-off, will not be considered as a neighbor interactor. In accordance with the fact that the median distance between two interacting proteins in a protein-protein network is not greater than 5, the parameter d, which has a negligible effect, is set to 2, as used in (Jia, P. *et al.*, 2011). The parameter r, instead has a considerable effect on the result, it obstructs restriction on the score of the module. Thus, during the module expanding process, if r is small,

a loose restriction is imposed in such way that unrelated nodes with lower z_i scores might be included, but if r is large, a strict restriction is imposed and only those nodes with very high z_i scores, and thus with very low p-values, would be included. In this case, some informative nodes with moderate p-values may be missed.

The searching strategy is performed using a greedy algorithm for iterative searching, in which each node in the network is considered as a seed. The follows steps are achieved in order to find dense modules (Jia, P. *et al.*, 2011):

Algorithm 3.1 Dense module searching (DMS) algorithm

1. A seed module is assigned. In the beginning, the seed module contains only the seed gene. Z_m is computed for the current seed module.
2. Identify neighborhood interactors, nodes with shortest path shorter than or equal to the predefined distance constraint d.
3. Examine the neighborhood interactors defined in Step (2) and find the genes generating the maximum increment of Z_m. Nodes will be added if the increment is greater than $Z_m \times r$, where r is the rate of proportion increment. That is, the expanded module has a score $Z_m + 1$ greater than $Z_m \times (1 + r)$.
4. Repeat Steps 1 - 3 until adding any neighborhood nodes cannot yield an increment that is greater than $Z_m \times r$.

Table 3.1 compares the two approaches, namely the dmGWAS and ancGWAS methods, emphasizing the technical improvements made in the new proposed method.

Table 3.1: Technical differences between dmGWAS and ancGWAS.

	dmGWAS	ancGWAS
Candidate sub-network searching	Yes	Yes
Annotation file & human PPI network	Yes	Yes
Pathway size adjustment	Yes	Yes
Gene weights	$p-$value	$p-$value, Ancestry proportion
p-value summary method	Fisher's, Simes, Smallest, FDR	Fisher's, Simes, Smallest, FDR
Edge weights	None	gene-LD
Subgraph enrichment	$p-$value	$p-$value, Ancestry proportion
Input data	List of SNP $p-$values	List of SNP $p-$values, SNP genotype data

3.1.2 Simulation of non-admixed pathway-based case-control population

To evaluate the performance of ancGWAS in detecting disease genes with weak genetic effects or strong epistatic effects that a single-marker-based testing approach from a standard GWA study can not detect, we simulated a pathway-based GWA study using PATHSIMU (Feng, Z. *et al.*, 2012). To do this, we used real genotype data from the CEU HapMap project to simulate 500 samples genotyped at $3,848,887$ markers. We simulated disease-predisposing genes (DPG) including *ATP5O*, *ATPIF1* and *BTG3* with disease-predisposing loci (DPL) *rs2834287*, *rs507238* and *rs2250305*, respectively, in the up-regulated aged mouse hypothalamus in NF-κB pathway. This pathway was randomly selected from a set of $1,047$ annotated pathways from the KEGG (http://www.genome.ad.jp/kegg/), BioCarta (http://www.biocarta.com/) and GeneAssist Pathway Atlas Databases (http://www.ambion.com/tools/pathway) as disease genes for a quantitative phenotype. The simulated disease pathway contains 37 additional genes, including *RPS12*, *MAP4*, *SNX2*, *NDUFB5*, *PRDX1*, *MAP2K1*, *AKR1A1*, *ANP32B*, *ATP5O*, *ATP5L*, *SEPT4*, *ATPIF1*, *ATP5C1*, *PPP1R7*, *ITPR1*, *BTG3*, *TPD52*, *CSF1R*, *C1QBP*, *PPA1*, *PSMA6*, *PSMA4*, *CCT6A*, *HK1*, *COX7A2*, *CTSS*, *PAFAH1B1*, *CASP6*, *PSMD14* *HEXB*, *ADCY9*, *TBCA*, *HSPE1*, *CLK1*, *ACTB*, *PREP* and *M6PR*.

The proportions of phenotypic variance explained by additive genetic effects of *ATP5O*, *ATPIF1* and *BTG3* were assigned 2%, 1% and 1%, respectively. We simulated interactive genetic effects between *ATP5O* and *ATPIF1*, *ATP5O* and *BTG3*, and *ATPIF1* and *BTG3*, explaining 0.5%, 0.3% and 0% of phenotypic variances, respectively. From the resulting simulated data, we then conducted the association analysis by applying EMMAX. The simulated GWA study data was subsequently analyzed using the ancGWAS and dmGWAS approaches.

3.1.3 Simulation of admixed case-control population

We used chromosome 1 with 116,413 autosomal SNPs from the HapMap3 project (International HapMap Consortium, 2010) populations, including CEU (Utah residents with ancestry from northern and western Europe), YRI (Yoruba in Ibadan, Nigeria), GIH (Gujarati Indians in Houston, Texas) and CHB (H an Chinese in Beijing, China) data. We independently expanded CEU, YRI, GIH and CHB to an additional 2,000 samples. To identify the occurrence of the admixture event (Chimusa, E. R. *et al.*, 2014), we sampled the haplotypes from YRI, CEU, GIH and CHB with a fixed probability (ancestral proportion) **60**%, 20%, 12% and 8%, respectively. The choice of ancestral proportion in simulating diploid admixed individuals was arbitrary. Following the sampling process above, the chromosomal segment of the ancestral population was copied to the genome of the admixed individual, and we recorded the locus-specific ancestry (the true ancestry) which served to assess ancGWAS. While simulating admixed individuals (Chimusa, E. R. *et al.*, 2014), we simulated four causal SNPs in which the risk allele has a higher frequency in YRI than in other ancestral populations. This leads to the selection of cases with higher than average YRI ancestry at the disease locus. The four causal SNPS include *rs2297977*, *rs841404*, *rs790633* and *rs6664119* with heterozgyote risks $R = 1.5$, 2, 1.5 and 2 and homozygote risks $R = 2.25$, 4, 2.25 and 2.25 (causal models); and the null mode risk alleles set to $R = 1$, 0, 1 and 0 at each SNP, respectively. For the null model, we chose random subsets of 1,000 cases and 1,000 controls. For causal models, we chose a random subset of 1,000 controls, and then chose 1,000 cases from the remaining samples so that samples with $0:1:2$ reference alleles have relative probabilities $1:R{:}R^2$ of being selected.

Our simulation yielded the genomes of 1000 cases and 1000 controls of mixed ancestry from YRI, CEU, GIH and CHB, with a total of 116,413 SNP markers. The simulated causal loci are on regions $1p31.3$ (*IL23R* gene) for SNPs $rs2297977$ and $rs841404$, and $1p34.2$ (*SLC2A1* gene) for SNPs $rs790633$ and $rs6664119$. Of note, *IL23R* and *SLC2A1* are interacting genes. We conducted standard GWA study analysis on the final simulation data set by applying EMMAX (Zhou, X. *et al.*, 2012), which accounts for both population stratification and hidden relatedness. To account for interacting disease SNPs and moderate risk that may not reach the intrinsic genome-wide significance cut-off of $P < 5.00 \times 10^{-08}$ in the standard GWA study above, we applied ancGWAS on the simulation GWA study result. We incorporated in ancGWAS the true locus-specific ancestry generated from the simulation above and the estimated locus-specific ancestries from LAMP-LD (Baran, Y. *et al.*, 2012) in assessing our methods for combining the locus-specific ancestries at both gene and sub-network levels.

3.2 Results and Discussion

3.2.1 Assessing ancGWAS on a simulated pathway-based association study

We evaluated ancGWAS using a pathway-based case-control simulated data set. The standard single-marker-based association analysis using EMMAX in Table 3.2, failed to identify (cutoff $p < 5\times10^{-08}$) our three simulated interactive disease-associated loci *rs2834287* ($p = 1.1\times10^{-4}$) in the *ATP5O* gene, *rs507238* ($p = 0.48$) in the *ATPIF1*, and *rs2250305* ($p = 0.19$) IN THE *BTG3* gene. Thus arose the opportunity and need to use a pathway-based approach to analyse the combined effect of all SNPs within a gene and genes within a pathway. To detect the simulated interactive disease genes with weak genetic effects in the up-regulated aged mouse hypothalamus pathway, we applied ancGWAS on the GWA study dataset and compared its performance to that of dmGWAS. After mapping SNPs to their closest genes, 23,726 and 23,498 genes were retained from ancGWAS and dmGWAS, respectively. This slight difference in number of mapped genes can be explained by slight differences in implementation of SNP-gene mapping methods in ancGWAS and dmGWAS, such as the cut-off distance upstream or downstream at which a SNP must be located with respect to a gene to be mapped the gene. In ancGWAS, the user can decide on the required boundary distance to a gene which a SNP must be located to be included in the analysis. For this particular analysis, a boundary distance of 20Kb upstream and downstream was considered. ancGWAS performs some additional checks to ensure that the chosen distance includes the majority of SNPs with significant or moderate signals, so that significant SNPs are not discarded from subsequent analysis, even though they are located in flanking regions or just outside the boundary cutoff.

Four possible methods have been implemented in ancGWAS for combining association signal at the gene level including the Fisher's combination, Sime's combination, FDR, and Smallest methods (See Section 2.1). A summary p-value was computed for each gene from SNPs within the gene using the smallest method. The results in Table 3.3 displays the top significant/moderate genes from the ancGWAS analysis where we combined the effect of multiple SNPs for a gene to refine the association signal, accounting for the difference in number of SNPs between different genes. To measure the proportion of false positives, we computed the q-value for each adjusted p-value taking into account the 23,726 genes in our analysis. These q-values, expected to have the same magnitude because of their adjusted p-values, indicate a low level of false positives. Interestingly, our DPGs have substantially increased their signal; *ATPIF1* ($p = 9.12 \times 10^{-03}$), *ATP5O* ($p = 9.99 \times 10^{-04}$) and *BTG3* ($p = 5.15 \times 10^{-03}$).

To construct the LD-weighted PPI network, three methods are available in ancGWAS; closestLD, ZscoreLD and maxLD. These three methods give similar results, therefore for simplicity of presentation we only report on the simulation result using the closestLD method. Pairwise correlations between all 23,726 genes were computed using the *closestLD* method, and an undi-

Table 3.2: Top genetic markers with moderate/significant p-values obtained from the single-marker-based association analysis using simulation of pathway-based GWA study data. Disease susceptibility SNPs are in bold.

SNP	Gene	CHR	Region	A_1	A_2	P-value
rs16861642	*CP*	3	q25.1	A	G	1.45×10^{-06}
rs10060036	*KIF2A*	5	q12.1	C	T	2.18×10^{-06}
rs953121	*DCC*	18	q21.3	G	T	2.65×10^{-06}
rs4912500	*EPHB3*	3	q27.1	G	TT	3.07×10^{-06}
rs920822	*RDX*	11	p15.4	C	G	3.15×10^{-06}
rs2167071	*ANGPT2*	8	p23.1	A	G	3.35×10^{-06}
rs1882314	*FBRSL1*	12	q24.33	A	G	3.68×10^{-06}
rs9391129	-	-	-	-	-	4.31×10^{-06}
rs12595323	*RFX7*	15	q21.3	A	G	4.62×10^{-06}
rs9449387	-	-	-	-	-	6.24×10^{-06}
rs7444781	*SLIT3*	5	q35.1	A	C	6.73×10^{-06}
rs11724257	-	4	q24	A	G	7.72×10^{-06}
rs4730481	*IMMP2L*	7	q31.1	A	C	7.99×10^{-06}
rs13006685	-	2	p25.2	A	G	9.58×10^{-06}
rs4709327	*RPL21*	6	q25.3	A	T	1.00×10^{-05}
rs2343542	*AACS*	12	q24.31	A	T	1.00×10^{-05}
rs3756187	*ACOX3*	4	p16.1	C	G	1.10×10^{-05}
rs10140097	*RAD51L1*	14	q24.1	C	T	1.11×10^{-05}
rs2053108	*CADM2*	3	p12.1	G	T	1.14×10^{-05}
rs2834287	***ATP5O**, ITSN1*	21	q22.11	C	T	1.1×10^{-04}
rs507238	***ATPIF1**, SESN2, DNAJC8, MED18*	1	p35.3	T	C	0.48
rs2250305	***BTG3**, CXADR, CXADRP2*	21	q21.1	A	G	0.19

Abbreviation: CHR, Chromosome; SNP, Single Nucleotide Polymorphism.

rected LD-weighted PPI network was then constructed composed of 10,756 genes with 31,942 interactions. A topological test was performed on the constructed LD-weighted network of 10,756 pair-wise gene-gene interactions and we assessed whether there is really an opportunity for using topological properties of the network as a factor for clustering. Figure 3.1a shows that this network exhibits a small world property, as well as scale-free topology (Figure 3.1b), that is, the spread of information in this network can be achieved through an average distance of only 3.9, and the loss of important nodes can lead to the disruption of the network. Interestingly, though the genes *ATPIF1*, *ATP5O* and *BTG3* are the DPGs, none of them appear to play a crucial role in this network when looking at their local topological properties. For instance, looking at the degree centrality of these genes and whether they are central genes or

Table 3.3: Top genes with moderate/significant p-values after combining single SNP p-values to represent gene p-values using simulation of pathway-based GWA study data. A correction using permutation was computed for each p-value as well as the significance of the adjustment. Disease susceptibility genes are in bold.

Gene	P**-value**	**Adjusted** p**-value**	Q**-value**
KIF2A	2.17×10^{-06}	0.0516	0.00010
FLJ37543	2.17×10^{-06}	0.0516	0.00010
EPHB3	3.07×10^{-06}	0.0729	0.00010
hCG1651160	3.14×10^{-06}	0.0746	0.00010
OR52B4	3.14×10^{-06}	0.0746	0.00010
RDX	3.14×10^{-06}	0.0746	0.00010
LOC100287015	3.34×10^{-06}	0.0794	0.00010
MCPH1	3.34×10^{-06}	0.0794	0.00010
ANGPT2	3.34×10^{-06}	0.0794	0.00010
GALNT9	3.67×10^{-06}	0.0872	0.00010
FBRSL1	3.67×10^{-06}	0.0872	0.00010
MIR585	6.73×10^{-06}	0.1596	0.00010
SLIT3	6.73×10^{-06}	0.1596	0.00010
IMMP2L	7.98×10^{-06}	0.1893	0.00010
LRRN3	7.98×10^{-06}	0.1893	0.00010
SOD2	1.01×10^{-05}	0.2372	0.00010
FNDC1	1.01×10^{-05}	0.2372	0.00010
RPL21	1.01×10^{-05}	0.2372	0.00010
ACOX3	1.10×10^{-05}	0.2613	0.00010
CYP51A1	1.13×10^{-05}	0.2699	0.00010
ATP5O	9.99×10^{-04}	0.9759	0.00015
BTG3	5.15×10^{-03}	0.9759	0.00022
ATPIF1	9.12×10^{-03}	0.9759	0.00028

not, we observe that $D_c = 1$ for *ATPIF1*, $D_c = 4$ for *ATP5O* and $D_c = 1$ for *BTG3*, and none of these genes is a central gene. Mining or approximating sets of genes they collaborate with may then be useful to take on these DPGs.

We applied the searching algorithm 2.1 implemented in ancGWAS (Subsection 2.1.3) on the LD-weighted PPI network of $10,756$ genes. First, we found all the hubs of the networks and successively, the betweenness centrality, the closeness centrality and the eigenvector centrality measures for each node were computed. We computed the cut-offs for each centrality measure, and the intersection of the top genes from each measure were considered to be the set of central nodes. A central node (central gene) and all its neighbors at distance (path step) $d = 1$ made up

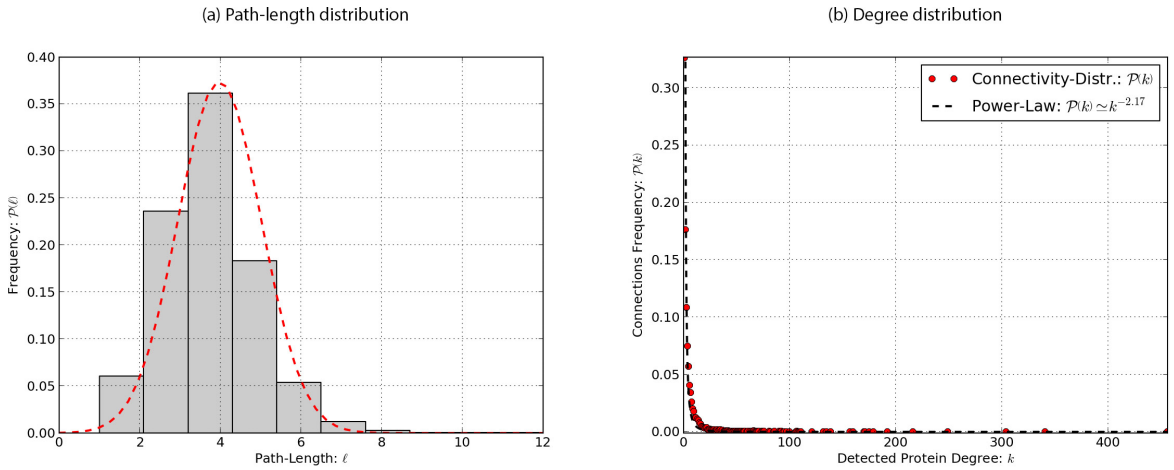

Figure 3.1: Topological analysis of LD-weighted PPI network for the pathway-based case-control simulated data set. (a) Distribution of shortest path lengths between reachable pair-wise protein functional interactions. This suggests that the transmission of biological information from a given gene to others is achieved through only a few steps. **(b)** Connectivity distribution of detected k functional links per protein, plotted as a function of frequency $\mathcal{P}(k)$. This suggests that most of the genes in this network have few interacting partners but some have many partners.

a module. We assessed the significance of each sub-network using the Stouffer-Liptak method. A total of 793 sub-networks (modules) were generated from ancGWAS. Table 3.4 lists the 20 top modules ranked by their corresponding p-values. Only hubs of modules are listed here for ease of presentation. For each module, we performed a pathway enrichment analysis to fully characterize generated sub-networks based on $1,047$ annotated pathways collected and curated from several public pathway databases (see Subsection 2.1.7).

Using the collected pathways, we compared each modules genes to those of collected pathways, and we computed the z-score for each pathway, and reported the number of genes overlapping between the module and the pathway with the highest z-score. We observed that the first module (*HSPA5*, MS = 70 genes, $p = 0.545$) has also shown the highest similarity score of ZP = 12.20, overlapping with the *ErbB receptor signaling network* for MOG = 32 different genes. ancGWAS incorporates prior knowledge on the disease genes, i.e. previously identified genes for the disease under study, and uses this information for mining for sub-networks of genes associated with the disease. For this simulation, a few genes were chosen to represent genes previously identified for the simulated disease, based on some criteria, including genes interacting with the simulated genes, in such a way that we can use these known genes to track unknown genes in some enriched sub-networks. Therefore, genes *PLOD2* ($p = 3.7 \times 10^{-04}$), *PLOD3* ($p = 4.2 \times 10^{-03}$), *RDH11* ($p = 6.7 \times 10^{-03}$), *UBQLN4* ($p = 6 \times 10^{-03}$), *PIK3CA* ($p = 0.024$), *CUL2* ($p = 4.7 \times 10^{-04}$), and *E2F1* ($p = 0.036$) were set as known disease genes. Interestingly, all the 20 top modules resulting contained 6 of these known genes (KG) (see Table 3.4), suggesting that integrating prior knowledge on known disease genes may significantly

Table 3.4: Top 20 sub-networks and related pathway enrichment results from ancGWAS using a simulated pathway-based GWA study data. For each pathway, the z-score (ZP) is reported representing the level of its similarity with the related module, which is also a function of the size of the module. Different scores of overlaps between the module, the pathway and the set of known genes are also reported.

Hub	P	Q	**MS**	**Pathway**	**ZP**	**MOG**	**MKG**	**PKG**
HSPA5	0.545	0.346	70	ErbB receptor signaling network	12.20	32	7	2
COMMD1	0.546	0.347	81	ErbB receptor signaling network	5.63	18	7	2
COPS6	0.547	0.347	143	Sphingosine 1-phosphate (S1P) pathway	5.01	26	7	2
KRT14	0.548	0.349	81	ErbB receptor signaling network	7.82	23	7	2
DCUN1D1	0.550	0.352	134	ErbB receptor signaling network	6.09	26	7	2
RPL19	0.566	0.371	76	ErbB receptor signaling network	6.66	18	7	2
CDKN2A	0.572	0.378	107	Sphingosine 1-phosphate (S1P) pathway	7.96	30	7	2
ELAVL1	0.575	0.382	525	Proteoglycan syndecan-mediated signaling events	5.58	66	7	2
SQSTM1	0.576	0.383	90	Proteoglycan syndecan-mediated signaling events	2.09	13	7	2
CSNK2A1	0.576	0.383	280	TRAIL signaling pathway	7.41	56	7	2
GRB2	0.580	0.388	336	TRAIL signaling pathway	11.26	83	7	2
KEAP1	0.585	0.395	90	ErbB receptor signaling network	6.44	20	7	2
HIST1H1C	0.585	0.395	81	Thrombin/protease-activated receptor (PAR) pathway	6.78	20	7	2
UBXN7	0.586	0.396	82	ErbB receptor signaling network	5.71	17	7	2
HDAC4	0.587	0.397	221	Sphingosine 1-phosphate (S1P) pathway	7.41	45	7	2
HSPA1A	0.588	0.399	108	IL3-mediated signaling events	9.34	29	7	2
GSK3B	0.594	0.405	246	TRAIL signaling pathway	9.08	57	7	2
RAN	0.595	0.408	199	Proteoglycan syndecan-mediated signaling events	5.34	37	7	2
TRAF2	0.596	0.408	244	TRAIL signaling pathway	7.97	49	7	2
YWHAB	0.596	0.408	242	TRAIL signaling pathway	7.08	47	7	2

Abbreviation. P: p-value; Q: q-value; MS: Module size; ZP: Pathway z-score; MOG: Gene overlapping between the module and the pathway; MKG: Number of genes overlapping between the module and the set of arbitrarily chosen genes as known disease genes; PKG: Number of genes overlapping between the pathway and the set of arbitrarily chosen genes as known disease genes;

increase the power to identify disease-specific pathways. In addition, ancGWAS was able to successfully detect one of our simulated disease genes (*ATPIF1*) in all 20 top modules.

We also applied dmGWAS on the PPI network of $10,756$ genes, and $8,931$ sub-networks were generated from dmGWAS using the default parameter setting of $d = 2$ and $r = 0.1$ as recommended in the dmGWAS method (see Table A.3 in supplementary materials). We retained the

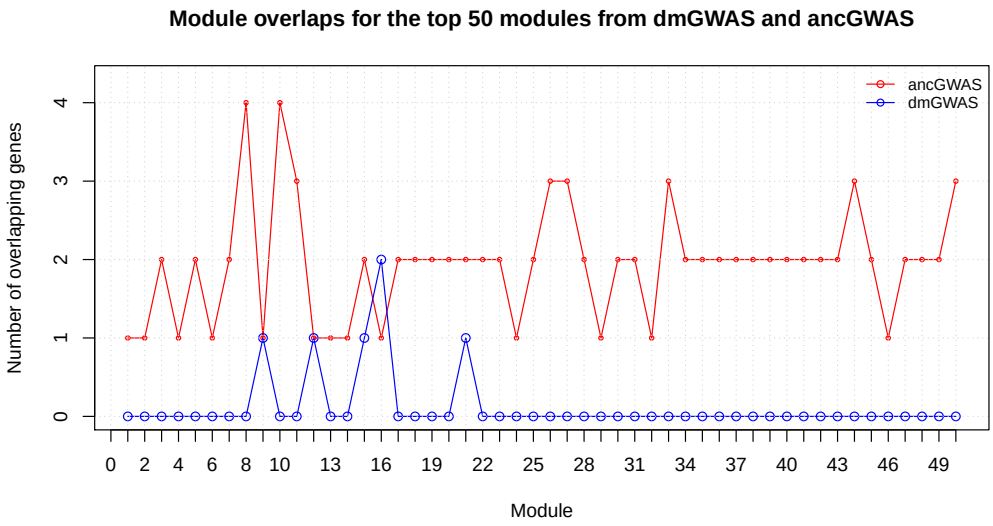

Figure 3.2: Number of overlapping genes between the top 50 sub-networks obtained from dmGWAS and ancGWAS, and the simulated disease pathway (*AGED MOUSE HYPOTH UP*), using the pathway-based simulated case-control data.

top 50 sub-networks from each approach (ancGWAS and dmGWAS), ranked according to their enrichment scores, and we subsequently compared the results from the two approaches. The simulated disease pathway was not found to be significantly enriched, however, we computed the number of genes overlapping between generated sub-networks with the up-regulated aged mouse hypothalamus pathway composed of 37 genes, including our DPGs, to assess whether each method was able to identify the simulated disease loci. We observed in Figure 3.2, that ancGWAS largely outperformed dmGWAS in approximating the simulated disease pathway in almost all the top 50 modules from both methods. This demonstrates that using topological properties of networks may significantly increase our chances of identifying disease associated sub-networks.

For each sub-network, we also computed the number of genes overlapping with our three simulated DPGs. Figure 3.3 shows that dmGWAS failed to detect any of the DPGs in any of its top 50 generated sub-networks. This may be attributable to not only the fact that it does not use network structural properties, but also that it does not integrate prior information of previously identified disease genes. As hypothesized earlier, understanding functional roles of interacting partners may facilitate the understanding of potential functional roles of as yet unannotated disease-related genes. In other words, known disease genes can be used to track unknown disease genes segregated in the same biological pathway, and this can be done using the interactions between these known and unknown genes. In this assessment, ancGWAS was able to successfully track gene *ATPIF1* because of its interaction with genes *UBQLN4* and *PIK3CA*, which were successfully identified by ancGWAS, and were considered to be known

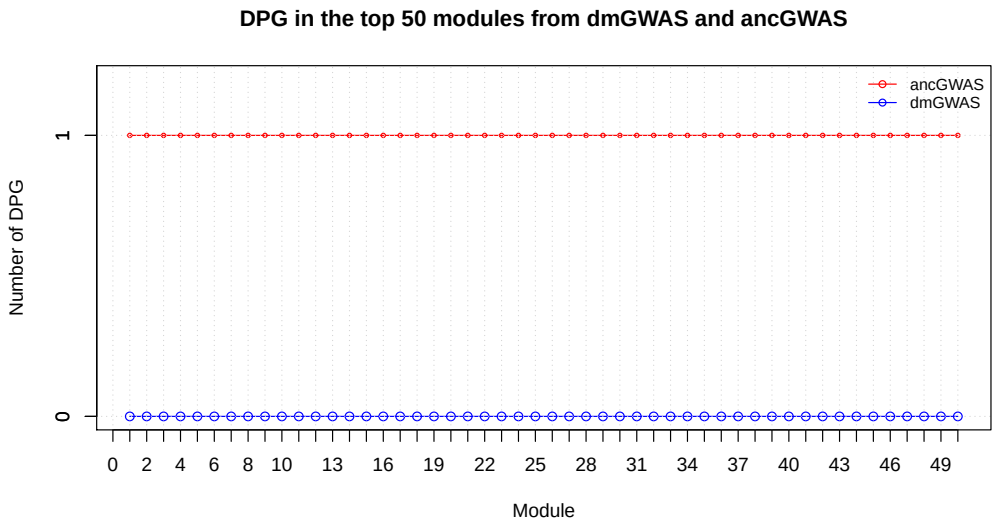

Figure 3.3: Number of overlapping genes between the top 50 sub-networks obtained from dmGWAS and ancGWAS, and the simulated disease-susceptibility genes (*ATP5O*, *ATPIF1* and *BTG3*), using the pathway-based simulated case-control data.

genes for the disease. The other 2 DPGs were not in the top 20 modules, which may be because we only used distance path length of one. In Chapter 4, we extend this path length.

Finally, for each pathway mapped to each sub-network from each method, we also computed the overlap with the simulated pathway, as well as with the three simulated DPGs (Figure 3.4). Again, the pathways enriched from ancGWAS contained more DPGs than those enriched from dmGWAS, even though a path length of two ($d = 2$) were used in dmGWAS.

Results in Figures 3.2, 3.3 and 3.4 show that ancGWAS largely outperformed dmGWAS, and demonstrated that the sub-networks obtained from ancGWAS better approximate the simulated disease pathway (*AGED MOUSE HYPOTH UP*), overlapping with disease-associated genes and with other genes within the disease pathway. This highlights the importance of characterizing susceptibility genes beyond standard GWA study analysis, which failed to detect the simulated disease gene signals.

3.2.2 Evaluating ancGWAS on a simulated disease in an admixed population

We evaluated ancGWAS using the simulated data of a 4-way admixed population with two disease loci at the *IL23R* gene in chromosomal region 1p31.3, and two other disease loci at the *SLC2A1* gene in chromosomal region 1p34.2 (see Section 3.1). We conducted the association

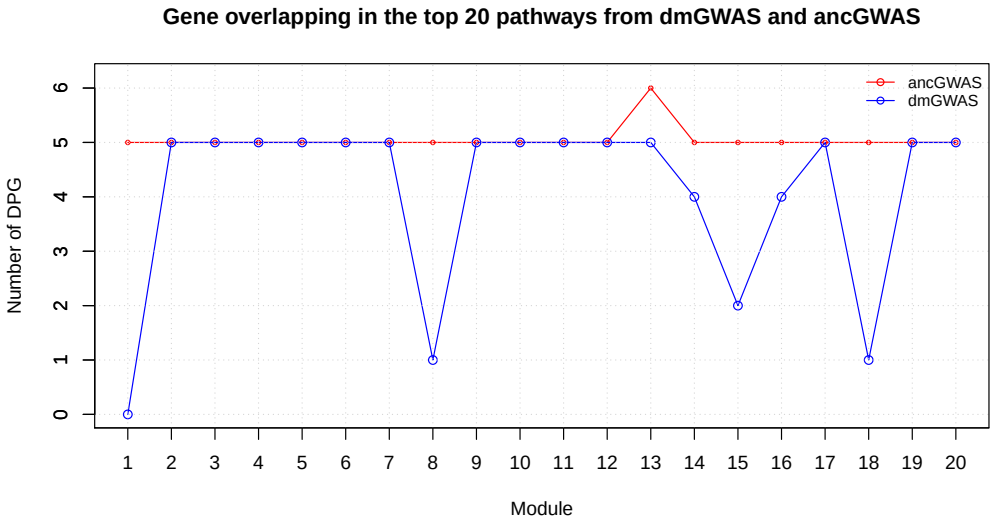

Figure 3.4: Number of overlapping genes between the top 20 pathways obtained from dmGWAS and ancGWAS module enrichment, and the simulated disease pathway (*AGED MOUSE HYPOTH UP*), using the pathway-based simulated case-control data.

analysis on this simulation data by applying EMMAX (Zhou, X. *et al.*, 2012), which accounts for both population stratification and hidden relatedness. Table 3.5 lists the top 23 most significant SNPs obtained from EMMAX, including the four simulated disease loci. Of note, EMMAX failed to identify as significant the simulated disease loci at SNPs *rs6664119* ($p = 0.48$), *rs2297977* ($p = 0.0357$), *rs841404* ($p = 1.51 \times 10^{-06}$) and *rs790633* ($p = 1.02 \times 10^{-05}$) and other related SNPs in LD with the simulated disease loci, including *rs841856* ($p = 0.65$) and *rs1385129* ($p = 0.0043$). To cover the moderate risk that did not reach the intrinsic genome-wide significance cut-off p-value of $< 5.00 \times 10^{-08}$ in this data (Table 3.5), we combine the effects of all SNPs in a particular gene, and the effects of genes at the pathway level using ancGWAS. Here, we first combined the GWA study data set and the true locus-specific ancestry obtained from the simulation of mixed ancestral populations.

The results in Table 3.6 display the top 20 moderate/significant genes from the ancGWAS analysis where we combined the effect of several SNPs for a gene to refine the association signal, accounting for the difference in number of SNPs between different genes (Yang, W. *et al.*, 2011). Interestingly, the simulated disease gene *SLC2A1* ($p = 6.60 \times 10^{-21}$), and other genes in *LD* with *SLC2A1* such as *SLC2A5* $\left(5.82 \times 10^{-25}\right)$, *SLC2A7* (1.05×10^{-23}), *C1orf210* (4.30×10^{-24}) and *FAM183A* (3.3×10^{-62}), which were on the boundary of genome-wide significance from the standard GWA study (Table 3.5), are now significant (Table 3.6) after combining effects of different SNPs within a gene. Some additional genes, including the simulated disease gene *C1orf141* (2.15×10^{-4}), that had weak signals from standard GWA analysis,

Table 3.5: 23 genetic markers with moderate/significant p-values obtained from the association analysis with simulated disease loci on the simulation data of the admixed population. Disease susceptibility SNPs are in bold.

SNP	Gene	CHR	A_1	A_2	P-value
rs841404	*SLC2A1*	1	T	C	1.51×10^{06}
rs2027286	*PCTK3*	1	T	C	3.68×10^{-06}
rs6695238	*EBNA1BP2, WDR65*	1	G	T	8.38×10^{-06}
rs790633	*IL23R, C1orf141*	1	T	C	1.02×10^{-05}
rs641350	*FAM183A*	1	G	T	1.10×10^{-05}
rs475093	*FAM183A*	1	G	C	1.10×10^{-05}
rs513009	*WDR65, EBNA1BP2, RNA5SP46*	1	A	G	1.32×10^{-05}
rs1308334	*FAM183A*	1	C	G	1.34×10^{-05}
rs558404	*EBNA1BP2*	1	T	C	1.72×10^{-05}
rs2453412	*EBNA1BP2, WDR65*	1	G	A	1.78×10^{-05}
rs11555249	*CDC20, ELOVL1, RP1-92O14*	1	G	A	1.98×10^{-05}
rs10890236	*FAM183A*	1	T	C	2.06×10^{-05}
rs12119303	*RP11-135J2*	1	G	T	2.79×10^{-05}
rs3120045	*BNA1BP2, WDR65*	1	G	A	2.84×10^{-05}
rs636969	*FAM183A*	1	T	G	3.56×10^{-05}
rs7550997	*CEP85*	1	G	A	3.94×10^{-05}
rs11555248	*CDC20, ELOVL1, RP1-92O14*	1	A	C	4.64×10^{-05}
rs768665	*EBNA1BP2, WDR65*	1	C	T	4.66×10^{-05}
rs6664119	*IL23R, C1orf141*	1	C	T	0.48
rs2297977	*SLC2A1*	1	G	T	0.036

are now moderate after combining signals from different SNPs within this gene. This result demonstrates the power of examining the combined effects of genes by detecting genetic signals beyond a single SNP.

We also tested for signals of unusual difference in deficiency/excess of ancestry under a null hypothesis. After combining the average locus-specific ancestry for each gene, we tested for bias of each observed average locus-specific ancestry and accuracy of the gene-specific ancestry estimates when combining local ancestry information. The results reported in Table 3.6 indicate no significant signal of unusual difference in deficiency/excess of ancestry, which is consistent with our simulation framework, which did not account for a model of differential ancestral allele frequency. In addition, this result can also be explained by the fact that the simulated time of the single admixture event was too recent to have an impact on deficiency/excess of ancestry in the simulated data. The gene level ancestry from mixed ancestral populations reported in Table 3.6 is proportional to the true ancestry proportion used to simulate the admixed population.

Table 3.6: Association analysis at the gene level on the simulation data of a 4-way admixed population. Top 20 genes with significant/moderate p-values obtained from the ancGWAS method of combined SNP association analysis with simulated disease on the simulation data of an admixed population. The table also displays ancestry-specific information from each ancestral population at the gene level, with its permutation p-value and the corresponding q-value in brackets.

Gene	P-value	CEU	CHB	GIH	YRI
FAM183A	3.3×10^{-62}	0.19 (0.66667, 0.00046)	0.187 (0.66667, 0.00046)	0.091 (0.66667, 0.00046)	0.532 (1.0, 0.00019)
WDR65	6.6×10^{-43}	0.19 (0.66667, 0.00045)	0.187 (0.66667, 0.00046)	0.091 (0.66667, 0.00045)	0.532 (0.94198, 0.00019)
ATP2B4	1.22×10^{-37}	0.254 (0.66667, 0.00044)	0.096 (0.66667, 0.00044)	0.040 (0.66667, 0.00044)	0.254 (0.67998, 0.00018)
ZC3H11A	6.39×10^{-37}	0.239 (0.63725, 0.00043)	0.115 (0.63725, 0.00043)	0.046 (0.63725, 0.00043)	0.591 (0.63725, 0.00018)
HHIPL2	1.7×10^{-35}	0.236 (0.66667, 0.00044)	0.132 (0.66667, 0.00044)	0.072 (0.66667, 0.00044)	0.558 (0.69056, 0.00018)
EBNA1BP2	1.86×10^{-35}	0.19 (0.66667, 0.00055)	0.187 (0.66667, 0.00055)	0.091 (0.66667, 0.00055)	0.531 (1.0, 0.00023)
LYPLAL1	4.97×10^{-35}	0.231 (0.66667, 0.00045)	0.130 (0.66667, 0.00045)	0.069 (0.66667, 0.00045)	0.572 (0.92483, 0.00019)
DUSP10	6.87×10^{-32}	0.236 (0.42349, 0.00037)	0.131 (0.42349, 0.00037)	0.073 (0.42349, 0.00037)	0.559 (0.42349, 0.00015)
CTTNBP2NL	1.03×10^{-28}	0.140 (0.66667, 0.00044)	0.145 (0.66667, 0.00044)	0.088 (0.66667, 0.00044)	0.624 (0.73621, 0.00018)
TMEM125	5.18×10^{-28}	0.19 (0.66667, 0.00052)	0.188 (0.66667, 0.00052)	0.090 (0.66667, 0.00052)	0.533 (1.0, 0.00021)
SLC2A5	5.82×10^{-25}	0.238 (0.66667, 0.00046)	0.137 (0.66667, 0.00046)	0.057 (0.66667, 0.00046)	0.568 (1.0, 0.00019)
C1orf210	4.30×10^{-24}	0.19 (0.66667, 0.000530)	0.188 (0.66667, 0.000530)	0.088 (0.66667, 0.000530)	0.533 (1.0, 0.00022)
SLC2A7	1.05×10^{-23}	0.237 (0.66667, 0.00047)	0.138 (0.66667, 0.00047)	0.057 (0.66667, 0.00047)	0.568 (.0, 0.00019)
GYG1	1×10^{-35}	0.239 (0.66667, 0.00047)	0.146 (0.66667, 0.00048)	0.062 (0.66667, 0.00047)	0.555 (1.0, 0.00019)
TIE1	1.84×10^{-23}	0.19 (0.66667, 0.00056)	0.188 (0.66667, 0.00056)	0.088 (0.66667, 0.00056)	0.533 (1.0, 0.00023)
ABCB10	4.46×10^{-23}	0.210 (0.66667, 0.00121)	0.14 (0.66667, 0.00121)	0.093 (0.66667, 0.00121)	0.557 (1.0, 0.0005)
ETV3	3.81×10^{-22}	0.28 (0.66667, 0.00047)	0.131 (0.66667, 0.00047)	0.038 (0.66667, 0.00047)	0.550 (1.0, 0.00019)
SLC2A1	6.60×10^{-21}	0.19 (0.38335, 0.00036)	0.187 (0.38335, 0.00036)	0.091 (0.38335, 0.00036)	0.531 (0.38335, 0.00015)
CEP85	3.82×10^{-5}	0.256 (0.66667, 0.00071)	0.097 (0.66667, 0.00071)	0.071 (0.66667, 0.00071)	0.576 (1.0, 0.00029)
C1orf141	2.15×10^{-4}	0.271 (0.66667, 0.00056)	0.131 (0.66667, 0.00056)	0.057 (0.66667, 0.00056)	0.541 (1.0, 0.00023)

CEU: Utah residents with ancestry from northern and western Europe; YRI: Yoruba in Ibadan, Nigeria; GIH: Gujarati Indians in Houston, Texas; CHB: Han Chinese in Beijing, China.

To benefit from fully characterizing susceptibility genes and determining the genetic structure of the simulated disease, we conducted sub-network association analysis using ancGWAS (see Section 2.1), using the closestLD method to construct the LD-weighted network. To this end, we used the PPI dataset from the PINA (http://cbg.garvan.unsw.edu.au/pina/interactome.stat.do) database containing 106, 573 human protein-protein interactions, reassembled and curated from different databases including IntAct, MINT, BioGRID, DIP, HPRD and MIPS/MPact.

A topological test was performed on the constructed LD-weighted network of 1, 742 pair-wise gene-gene interactions and we assessed whether there is really an opportunity for using topological properties of the network as a factor for clustering. Figure 3.5 shows that the network exhibits scale-free topology, meaning that the degree distribution of genes approximates a power law $P(k) = k^{-\gamma}$, where $\gamma \approx 2.70$ is the degree exponent obtained by fitting the model using the least-square approach. This indicates that most genes have few interacting partners but some have many and are crucial for the robustness of the network. Figure 3.5 shows that the network has a small world property, suggesting that the spread of information in the network is achieved through an average of 7.01 steps, which corresponds to the average shortest paths in the network. To determine whether we can use topological properties to break down our network into sub-networks, we ran the searching algorithm (see workflow 2.1). First, we found all the hubs of the networks and successively, the betweenness centrality, the closeness centrality and the eigenvector centrality measures for each node were computed. We computed the cut-offs for each centrality measure, and the intersection of the top genes from each measure was considered to be the set of central nodes. We assessed the significance of each of the top 20 sub-networks using the Stouffer-Liptak method, accounting for spatial correlations among SNPs within a gene or genes within a given sub-network.

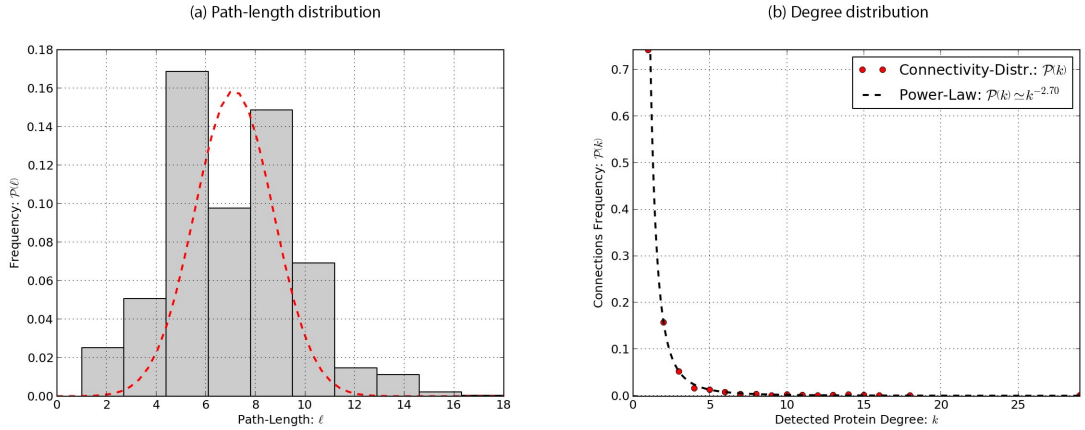

Figure 3.5: Topological analysis of the 4-way admixed population LD-weighted PPI network. **(a)** Distribution of shortest path lengths between reachable pair-wise protein functional interactions. **(b)** Connectivity distribution of detected k functional links per protein, plotted as a function of frequency $\mathcal{P}(k)$.

Table 3.7: Association analysis at the sub-network level on the simulation data of a 4-way admixed population. Top 20 sub-networks (only module hubs are displayed) with significant/moderate p-values obtained from the ancGWAS method of combined SNP association analysis with simulated disease on the simulation data of an admixed population. The table also displays ancestry-specific information from each ancestral population at the gene level, with its permutation p-value and the corresponding q-value.

Module Hub	Liptak	Q	YRI	GIH	CHB	CEU
TRAF3IP3	8.26e-31	0.00.0	0.553 (0.25, 0.02)	0.076 (0.25, 0.02)	0.132 (0.25, 0.02)	0.231 (0.25, 0.02)
CTTNBP2NL	8.26e-31	0.00.0	0.576 (0.25, 0.02)	0.053 (0.25, 0.02)	0.15 (0.25, 0.02)	0.228 (0.25, 0.02)
RPSA	9.82e-19	0.00.0	0.55 (0.25, 0.034)	0.06 (0.25, 0.034)	0.146 (0.25, 0.034)	0.239 (0.25, 0.034)
SLC2A5	9.82e-19	0.00.0	0.57 (0.25, 0.034)	0.068 (0.25, 0.034)	0.134 (0.25, 0.034)	0.232 (0.25, 0.034)
SRSF10	9.10e-17	4.94e-09	0.571 (0.2, 0.005)	0.056 (0.2, 0.005)	0.141 (0.2, 0.005)	0.235 (0.2, 0.005)
LEPR	3.62e-14	9.17e-08	0.573 (0.25, 0.027)	0.054 (0.25, 0.027)	0.128 (0.25, 0.027)	0.248 (0.25, 0.027)
IPO13	3.08e-12	8.69e-07	0.609 (0.102, 0.003)	0.045 (0.102, 0.003)	0.117 (0.102, 0.003)	0.236 (0.102, 0.003)
SNRPE	8.82e-12	1.48e-06	0.582 (0.2, 0.005)	0.054 (0.2, 0.005)	0.131 (0.2, 0.005)	0.244 (0.2, 0.005)
CTSD	9.03e-12	1.49e-06	0.575 (0.102, 0.003)	0.038 (0.102, 0.003)	0.127 (0.102, 0.003)	0.267 (0.102, 0.003)
PSMA5	9.03e-12	3.11e-06	0.574 (0.102, 0.003)	0.054 (0.102, 0.003)	0.131 (0.102, 0.003)	0.252 (0.102, 0.003)
S100A14	9.03e-12	3.11e-06	0.574 (0.102, 0.003)	0.054 (0.102, 0.003)	0.131 (0.102, 0.003)	0.252 (0.102, 0.003)
PSMB4	9.03e-12	3.11e-06	0.574 (0.102, 0.003)	0.054 (0.102, 0.003)	0.131 (0.102, 0.003)	0.252 (0.102, 0.003)
RPL31	9.33e-11	4.90e-06	0.562 (0.102, 0.003)	0.061 (0.102, 0.003)	0.129 (0.102, 0.003)	0.242 (0.102, 0.003)
CGN	3.58e-10	9.72e-06	0.59 (0.2, 0.005)	0.055 (0.2, 0.005)	0.124 (0.2, 0.005)	0.24 (0.2, 0.005)
STXBP3	4.96e-10	1.14e-05	0.61 (0.2, 0.005)	0.047 (0.2, 0.005)	0.118 (0.2, 0.005)	0.233 (0.2, 0.005)
NID1	1.06e-09	1.68e-05	0.62 (0.102, 0.003)	0.066 (0.102, 0.003)	0.142 (0.102, 0.003)	0.18 (0.102, 0.003)
RPAP2	2.62e-09	2.67e-05	0.609 (0.092, 0.003)	0.045 (0.092, 0.003)	0.117 (0.092, 0.003)	0.237 (0.092, 0.003)
SFN	1.26e-08	5.96e-05	0.609 (0.2, 0.005)	0.046 (0.2, 0.005)	0.117 (0.2, 0.005)	0.239 (0.2, 0.005)
LAMC1	2.07e-08	7.67e-05	0.623 (0.25, 0.008)	0.08 (0.25, 0.008)	0.145 (0.25, 0.008)	0.163 (0.25, 0.008)
APOA2	2.58e-08	8.59e-05	0.556 (0.25, 0.034)	0.056 (0.25, 0.034)	0.139 (0.25, 0.034)	0.25 (0.25, 0.034)

Q: q-value of the Liptak test statistics;

Table 3.7 reports the ancestral proportions of each sub-network for each ancestry component, which is consistent with ancestral proportions used to simulate this population data. A total of 382 modules were generated from ancGWAS, and ranked according their Liptak statistics. For each ancestry, the adjusted Wilcoxon signed-rank statistic p-value and the corresponding type 1 error level (q-value) are reported. No evidence of difference in deficiency/excess of ancestry is shown for each those top 20 sub-networks, revealing that gene-specific ancestries from cases and controls in this population are the same. Importantly, when applying the enrichment analysis implemented in ancGWAS to the top 20 sub-networks (Table 3.8), the annotations for pathway/process of the 20 top sub-networks are clustered in signaling pathways, which include the pathway *Proteoglycan syndecan-mediated signaling events*, containing many of our known genes, and more particularly our simulated disease gene *SLC2A1*. The result in Table 3.8 highlights the benefit of fully characterizing susceptible genes beyond the standard GWA study approach. A strong convergence of almost all the pathways associated to modules generated by ancGWAS was observed, suggesting a possible convergence of all these pathways into a single network.

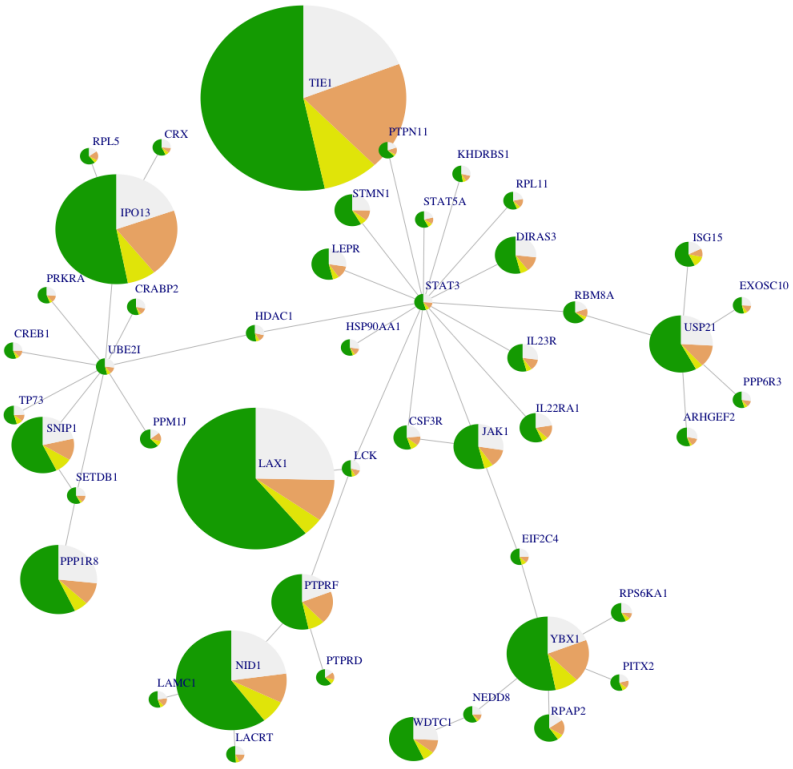

Figure 3.6: Central network of 47 genes for the 4-way simulated data from ancGWAS. The size of a node denotes its significance with size increasing with significance. Each gene is represented with its ancestral proportions: green for YRI; yellow for GIH; sandy-brown for CHB, white-smoke for CEU.

Table 3.8: Enrichment analysis of sub-netwokrs using ancGWAS on the simulation data of a 4-way admixed population.
Top 20 sub-networks represented by their hubs generated using ancGWAS, ranked according to their Liptak's combined p-value test statistic. See Table A.1 in supplementary materials for a more detailed list of sub-networks and corresponding genes. For each sub-network (module), the most related pathway is reported with respect to the significance of overlap between the sub-network and the human pathway. ancGWAS was able to identify enriched sub-network harboring one of our simulated genes (*IL23R*), and one of the genes (*UBE2I*) arbitrarily considered as previously identified to be associated with the simulated disease in ancGWAS.

Module Hub	MS	Pathway	ZP	MOG	MKG	PKG	PDPG
TRAF3IP3	6	Apoptotic execution phase	112.67	1	-	-	-
CTTNBP2NL	6	Plasma membrane estrogen receptor signaling	2.99	1	-	7	*SLC2A1*
RPSA	3	Proteoglycan syndecan-mediated signaling events	2.82	1	-	7	*SLC2A1*
SLC2A5	3	Proteoglycan syndecan-mediated signaling events	2.82	1	-	7	*SLC2A1*
SRSF10	13	ATR signaling pathway	22.51	1	-	-	-
LEPR	3	Signaling events mediated by PTP1B	104.92	1	-	2	-
IPO13	26	Proteoglycan syndecan-mediated signaling events	2.39	6	*UBE2I*	7	*SLC2A1*
SNRPE	17	Proteoglycan syndecan-mediated signaling events	1.69	3	-	7	*SLC2A1*
CTSD	32	Immune System	14.48	6	-	3	-
PSMA5	29	Immune System	12.62	5	-	3	-
S100A14	29	Immune System	12.62	5	-	3	-
PSMB4	29	Immune System	12.62	5	-	3	-
RPL31	24	Proteoglycan syndecan-mediated signaling events	4.65	6	*UBE2I*	7	-
CGN	15	CDC42 signaling events	6.01	3	-	6	*SLC2A1, IL23R*
STXBP3	15	Proteoglycan syndecan-mediated signaling events	1.69	*3*	-	7	*SLC2A1*
NID1	25	Beta1 integrin cell surface interactions	8.44	13	*UBE2I*	7	*SLC2A1*
RPAP2	34	Proteoglycan syndecan-mediated signaling events	4.42	11	*UBE2I*	7	*SLC2A1*
SFN	15	ATR signaling pathway	28.27	2	-	-	-
LAMC1	11	Beta1 integrin cell surface interactions	6.32	5	-	7	*SLC2A1*
APOA2	2	Processing of Capped Intron-Containing Pre-mRNA	0.0	-	-	1	-

Abbreviation. MS: Module size; ZP: Pathway z-score; MOG: Gene overlapping between the module and the pathway; MKG: Number of genes overlapping between the module and the set of arbitrary chosen genes as known disease genes; POG: Number of overlapping genes between the pathway and the known genes; PDPG: Number of DPGs in the pathway.

Taking advantage of overlapping genes among these sub-networks, we looked for a central sub-network from the intersection of all the sub-networks. The central sub-network in Figure 3.6 represents the most important sub-network for this simulated data set, and it highlights important features of why pathway-based or network-based approaches may be crucial as a complementary approach to the single-marker-based approach in GWA studies. The central gene *STAT3*, though it doesn't show a significant association signal itself, is really important for the survival of this system, and its removal would result in a complete breakdown of this network. It can lead to fatal loss of substantial information in this sub-network, such as the loss of the simulated disease gene *IL23R*, and other nodes that hold this network together including genes *TIE1*, *ipo13*, *UBE21* and *YBX1* (Figure 3.6). This gene/protein plays a key role in many cellular processes such as cell growth and apoptosis. It has been associated with many diseases including hyper ige syndrome, and autosomal dominant hyper ige syndrome. In the IL-23 pathway for instance, STAT3 is required for IL-23-mediated IL-17 production in spontaneous arthritis animal model IL-1 receptor antagonist-deficient mice (Cho, M. *et al.*, 2006). In another experiment, IL-23 and IL-12 expression in tumor-infiltrating myeloid cells have been shown to be compartmentalized and oppositely regulated by Stat3, by conditional knock out of the Stat3 gene in the hematopoietic compartment (Kortylewski, M. *et al.*, 2009).

Taken together, through the simulation of a 4-way admixed population, ancGWAS demonstrated accuracy and the ability to elucidate the interactions between genes underlying the pathogenesis of complex diseases that were not detected in a standard GWA study analysis. It was also able to find ancestry-specific genes or sub-networks, and showed no evidence of no deficiency/excess differences in ancestry at the SNP, gene and pathway level.

We have shown, through simulation of a pathway-based association study and simulation of interactive disease loci in an admixed disease population, that ancGWAS holds promise for comprehensively examining interactions between genes underlying the pathogenesis of genetic diseases and also underlying ethnic differences. We demonstrated that ancGWAS outperformed an existing method, dmGWAS, by leveraging the correlation that exists among SNPs within genes and genes within pathways, and accounting for the topological structure of the network. Importantly, ancGWAS was able to recover weak and moderate association signals at the gene and pathway level. It also refined the weak signals of simulated disease SNPs, which failed in the single-marker-based testing approach commonly used in standard GWAS, to identify significant and enriched sub-networks associated with complex disease.

CHAPTER 4

Application of ancGWAS: Identification of enriched pathways for sporadic postmenopausal breast cancer

Breast cancer is a disease in which certain cells in the breast become abnormal and multiply without control or order to form a tumor. Like most complex diseases, it is likely to be a disruption of a mechanism of several interacting genes rather than a result of a single gene, and as in most cancers, it represents a heterogeneous collection of distinct diseases that arise as a consequence of carried somatic mutations acquired during tumorigenesis (Hanahan, D. *et al.*, 2000). The ability to dissect this heterogeneity with respect to the mechanism leading to the disease, as well as with respect to the nature of the disease itself, is crucial for the understanding of the disease etiology. This can critically help understanding the significance of the genome alterations in breast cancer, and therefore developing more effective therapeutic strategies for personalized medicine, by identifying groups or subpopulations of individual patients of particular characteristics who are likely to respond positively or negatively to a particular therapeutic strategy (Chuang, H. Y. *et al.*, 2007). Though most cases of breast cancer are not inherited, relatively low heritability has been established, estimated to less than 30% of the genetic component involved in the pathogenesis of breast cancer (Chang, E. *et al.*, 2014). For instance, mutations in the *BRCA1* and *BRCA2* genes (major genes related to hereditary breast cancer), are inherited in an autosomal dominant pattern, so that one copy of each gene in each cell can sufficiently increase a person's risk of developing breast cancer (National Institute of Health, 2014).

Over the years, several different genetic methodologies have been employed to identify multiple loci variations in population frequency that can confer susceptibility, relative risk or play a potential functional role in breast cancer (Table 4.1). These range from familial linkage and positional cloning studies, which led to the identification of the *BRCA1* and *BRCA2* genes

Table 4.1: Known breast cancer susceptibility genes and regions showing the mapping method used to infer the association. A more complete list is provided in Table A.2 in supplementary materials. Adapted from (Collins, A. *et al.*, 2011).

Known gene	Location	Mapped by	All. Freq.	Known/possible function
BRCA1	17q21	Linkage	Rare	DNA repair/genome stability
BRCA2	13q13.1	Linkage	Rare	Recombinational repair
TP53	17q12.1	Linkage	Rare	Li–Fraumeni syndrome, apoptosis
ATM	11q22.3	Candidate	Rare	DNA repair
BRIPI	17q23.2	Candidate	Rare	DNA repair, associated with BRCA1
CHEK2	22q12.1	Candidate	Rare	DNA repair/cell cycle
PALB3	16p12.2	Candidate	Rare	Associated with BRCA2
RAD51C	17q22	Candidate	Rare	Homologous recombination repair
PTEN	10q23.3	Linkage	Rare	Cowden disease, cell signaling
STK11 (LKBI)	19p13.3	Linkage	Rare	Peutz–Jeghers syndrome, cell cycle arrest
CDHI	16q22.1	Linkage	Common	intercellular adhesion: lobular BC
FGFR2	10q26	GWAS	Common	Fibroblast growth factor receptor
TOX3 (TNRC9)	16q12	GWAS	Common	Chromatin structure/cell cycle
MAP3KI	5q11.2	GWAS	Common	Cellular response to growth factors
LSPI	11p15.5	GWAS	Common	Neutrophil motility
CASP8	2q33	GWAS	Common	Apoptosis
SLC4A7/NEK10	3p24.1	GWAS	Common	Cell cycle control?
NOTCH2/FCGR1B	1p11.2	GWAS	Common	Signaling/immune response?
RAD5ILI	14q24.1	GWAS	Common	Homologous recombination repair?
CDKN2A/CDKN2B	9p21	GWAS	Common	Cyclin-dependent kinase inhibitors?
MYEOV/CCNDL	11q13	GWAS	Common	Cell cycle control/fibroblast growth factors?
ZNF365	10q21.2	GWAS	Common	Zinc finger protein gene
ANKRD16/FBX018	10p15.1	GWAS	Common	Helicase
ZMIZI	10q22.3	GWAS	Common	Regulates transcription factors?

Abbreviation. All. Freq.: Allele frequency; BC: breast cancer.
Notes. ? refers to 'possible'/'uncertain' gene or function in the breast cancer context. Candidate refers to mapping by candidate resequencing.

that belong to the DNA repair mechanism in the cells and account for a substantial proportion of early-onset breast cancer, and the tumor suppressor *TP53* and *PTEN* genes that participate in processes related to cell cycle control and cell proliferation (Ideker, T. *et al.*, 2002; Collins, A. *et al.*, 2011), to candidate gene association studies that have led to the identification of additional rare genetic variants with relatively moderate risks. These include germline mutations in the ataxia-telangiectasia (*ATM*) gene (with apparently higher risks below the age of 50 years) and the partner and localizer of the *BRCA2* (*PALB2*) gene, which interacts with *BRCA2*, and in which mono-allelic mutations are involved in familial breast cancer (Collins, A. *et al.*, 2011). More recently, GWA studies have led to the discovery of several novel common

low penetrance risk alleles such as the well-established *FGFR2* gene in which common susceptibility variants are located in intron 2 (Hunter, D. J. *et al.*, 2007). Several other studies have looked at more mechanistic possibilities taking into account interaction effects between genes ($G \times G$), and between genes and the environment ($G \times E$). For example, Briollais et *al* in (Briollais, L. *et al.*, 2007) examined SNP-SNP interactions among 19 SNPs from 18 key genes involved in major cancer pathways in a sample of 398 breast cancer cases and 372 controls from Ontario, and identified several simple (two-way) and complex (multi-way) SNP-SNP interactions associated with breast cancer, including an interaction between *XPD* and *IL10* genes as the most significant two-way interaction. Many other studies of this nature have been carried out over the years, inspecting the link between groups of genes and breast cancer, resulting in the identification of many targeted pathways for breast cancer, including the estrogen signaling pathway, the phosphoinositide-3 kinase (*PI3K*) pathway, the HER2 signaling pathway and the insulin-like growth factor (*IGF*) signaling pathway (Chang, J. T. *et al.*, 2009; Baselga, J., 2011; Ideker, T. *et al.*, 2002), of which many are currently actively used for targeted therapies for breast cancer treatment (National Cancer Institute, 2014).

Here, we applied the new proposed network-based approach, ancGWAS, to the breast cancer GWAS dataset of the Cancer Genetic Markers of Susceptibility (http://cgems.cancer.gov/) project of the National Cancer Institute (NCI, http://www.cancer.gov/). A recent study in (Hunter, D. J. *et al.*, 2007) has identified significant association in intron 2 in the *FGFR2* gene with breast cancer susceptibility. We first describe the data used in this analysis, and then present the result obtained through ancGWAS.

4.1 Materials and Methods

We obtained the stage I case-control breast cancer data (Hunter, D. J. *et al.*, 2007), from the Cancer Genetic Markers of Susceptibility (CGEMS) project. The data set was genotyped on an Illumina HumanHap 550 array, and the samples included $1,145$ postmenopausal women of European ancestry with invasive breast cancer and $1,142$ controls, nested within the prospective Nurses Health Study cohort. Data quality control was performed, excluding all unmapped SNPs. $528,169$ SNPs were finally included in this study. After evaluating the extent of substructure in the data set (separating cases and controls as distinct groups), we examined whether stratification can be accounted for in the GWAS. Both population stratification and hidden relatedness were taken into account by applying the mixed model approach in EMMAX on the dataset. In addition, we conducted an imputation (Howie, B. N. *et al.*, 2009), by only incorporating genotypes in the chromosomal region $10q26$ that harbours the *FGFR2* gene, based on the $1,000$ Genomes project populations, to ensure that the majority of SNPs in that gene are present in our analysis.

528, 169 SNPs genotyped from 1, 145 cases and 1, 142 controls were tested for association with breast cancer, including the additional imputed SNPs. Though the association analysis did not yield any genome-wide significant signal, the *FGFR2* gene was highly associated with breast cancer susceptibility (Table 4.2). To account for possible interacting breast cancer disease SNPs and moderate risk that could not reach the genome-wide significance cut-off in the standard GWAS, and to investigate the joint effect of interacting groups of genes that might be associated with breast cancer, we applied ancGWAS to the resulting GWAS data set containing 528, 169 SNPs with their corresponding p-values.

4.2 Results and discussion

We applied the method presented in Chapter 2, and implemented in ancGWAS. The results in this section are reported as a series of distinct stage (stepwise process) as presented in the work-flow in Figure 2.1, highlighting the approach and results at different stage of the analysis performed in ancGWAS.

After mapping SNPs to genes at a boundary distance of 20Kb upstream and downstream, 24, 403 genes were retained in ancGWAS, representing 319, 137 of the 528, 169 SNPs (60%) that were originally genotyped in this GWAS. Summary p-values were computed for genes from SNPs using the smallest method with a cut-off of 0.05, also accounting for the difference in number of SNPs between different genes. Table 4.3 displays the top significant/moderate genes from the first step in the ancGWAS analysis, combining effects of multiple SNPs for a gene to refine the association signal. Looking at some known breast cancer genes from previous association studies, we observed that the majority of these genes still did not have significant association signal in this dataset. Nevertheless some genes such as *HAS2-AS1* ($p = 3.46 \times 10^{-07}$), *MYCL1* ($p = 2.73 \times 10^{-06}$), and *RAD51C* ($p = 0.00299$) have their association signal increased after combining p-values from SNPs. However, many of them have still shown very weak association signals, the *FGFR2* gene ($p = 0.03702$) signal decreased, and at the extreme, no signal of association was found for *BRCA1* ($p = 1.0$), one of high-penetrance breast cancer genes. Combining association signals from SNPs for a gene does not necessarily increase the significance of the signal. For instance, the *FGFR2* gene, which has the highest signal from the single-marker-based association analysis ($p = 2.5 \times 10^{-06}$), now has its signal decreased ($p = 0.03$ and adjusted $p = 0.14$). 138 SNPs are associated with the *FGFR2* gene, 90% of which have weak signals greater than or equal to 0.1, which causes the decreasing of the signal of this gene.

Pairwise correlation between all these 24, 403 genes were computed using the *closestLD* method, and an undirected LD-weighted PPI network was then constructed composed of 10, 839 genes with 34, 665 interactions. The topological analysis of this LD-weighted PPI network reveals in Figure 4.1a that this network exhibits a small world property, as well as a scale-free topology

Table 4.2: Top genetic markers with moderate/significant p-values obtained from the association analysis using $1,145$ postmenopausal women of European ancestry with invasive breast cancer and $1,142$ controls, genotyped at $528,169$ SNPs.

SNP	**Gene**	**CHR**	**Region**	A_1	A_2	P**-value**
rs10510126	*FGFR2*	10	q26.13	C	T	$2.56e-06$
rs12505080	*KIAA1239*	4	p14	T	C	$7.98e-06$
rs17157903	*RELN*	7	q22.1	C	T	$1.02e-05$
rs1219648	*FGFR2*	10	q26.13	A	G	$1.34e-05$
rs7696175	*TLR1, TLR6*	4	p14	C	T	$1.40e-05$
rs2420946	*FGFR2*	10	q26.13	C	T	$1.73e-05$
rs6497337	*SYT17*	16	P12.3	G	A	$2.12e-05$
rs1250255	*FN1*	2	q35	A	G	$3.21e-05$
rs10804287	*EPHA4, MIR4268*	2	q35	G	A	$3.63e-05$
rs16943326	*PPM1E*	17	q22	C	T	$4.30e-05$
rs2981579	*FGFR2*	10	q26.13	C	T	$4.36e-05$
rs1538472	*KIF26B*	1	q44	T	C	$4.53e-05$
rs873811	*C21orf82, MRPS6*	21	q22.11	C	T	$5.50e-05$
rs4107736	*TMEM66, MIR3148*	8	p12	A	G	$5.53e-05$
rs11732323	*KIT, KDR*	4	q12	G	A	$5.77e-05$
rs2194225	*IL7R*	5	p13.2	A	G	$6.56e-05$
rs12792184	*OR8A1*	11	q24.2	T	C	$7.07e-05$
rs12682293	*MAP2K1P1*	8	p12	C	A	$7.12e-05$
rs4445554	*USP6NL*	10	p14	C	T	$7.16e-05$
rs4866929	*HCN1*	5	p12	A	G	$7.17e-05$
rs9953717	*CBLN2, SOCS6*	18	q22.3	C	T	$7.65e-05$

(Figure 4.1b), so the spread of information in this network can be achieved through an average distance of only 3.9. This shows a high connectivity in this network, with a number of important genes with many partners. The network average degree is 2, which means that every gene in this network has on average 2 partners.

We applied the searching algorithm from 2.1 implemented in ancGWAS on the LD-weighted PPI network of $10,839$ genes. After finding all the hubs of this network, we subsequently computed the betweenness, closeness and eigenvector centrality measures for each node. Cut-offs *BetOf*, *ClosOf*, and *DegOf* for each centrality measure were computed and applied to every node centrality score, resulting in *BetOf*$=$ 42033.38, *ClosOf* $=$ 0.2579, *DegOf*$=$ 6.288. The intersection of the set of genes for each centrality measure was considered to be the set of central nodes. Finally, for each central node (central gene), all its direct neighboring partners at distance (path step) $d = 1$ made up a module.

Table 4.3: Top genes with moderate/significant p-values obtained from the ancGWAS method of combined SNP association analysis using 1,145 breast cancer cases and 1,142 controls, genotyped at 528,169 SNPs. In *bold* are some genes identified in previous association studies as being involved in breast cancer.

Gene	P-values	AdjP GC	Gene	P-values	AdjP GC
VWA3B	3.08×10^{-13}	2.84×10^{-07}	*ASCC3*	9.31×10^{-08}	0.00017
MRPS30	7.33×10^{-12}	1.40×10^{-06}	*SLC4A3*	2.53×10^{-07}	0.00028
RRAGA	7.33×10^{-12}	1.40×10^{-06}	*ATG10*	3.43×10^{-07}	0.00033
PTCD3	1.35×10^{-11}	1.92×10^{-06}	***HAS2-AS1***	3.46×10^{-07}	0.00033
IMMT	3.57×10^{-10}	1.00×10^{-05}	***MYCL1***	2.73×10^{-06}	0.00095
SCARNA8	4.20×10^{-09}	3.51×10^{-05}	***SOD2***	0.00065	0.01644
TADA1L	4.26×10^{-09}	3.54×10^{-05}	***RAD51C***	0.00299	0.03869
CNGA3	4.89×10^{-09}	3.79×10^{-05}	***SLC4A7***	0.02171	0.10990
JRKL	1.01×10^{-08}	5.51×10^{-05}	***CCND1***	0.01044	0.07139
TANK	2.98×10^{-08}	9.52×10^{-05}	***BARD1***	0.02597	0.11694
TSPAN15	3.86×10^{-08}	0.00010	***MPO***	0.02732	0.12022
MMRN1	3.93×10^{-08}	0.00010	***FGFR2***	0.03702	0.14199
BMPR1B	4.02×10^{-08}	0.00011	***PTGS2***	0.04438	0.15690
C6orf10	5.74×10^{-08}	0.00013	***BRCA1***	1.0	1.0
KRT18	8.49×10^{-08}	0.00016			

At different stages of any research project, investigators need to choose which genes or proteins to investigate further experimentally and which to leave out for particular reasons. This is often done randomly or using a statistical method, and is referred to as "*gene prioritization*". The basic idea behind gene prioritization is that for a given phenotype with genetic heterogeneity, different trait-related genes should exhibit some similarities with one another based on a particular measure. In other words, if we assume a set of genes responsible for a given disease ("seed genes"), here referred to as "known disease genes" (KDG), then unknown disease genes can be detected through their similarities (interactions in the case of PPIs) to the seeds (Bao, S. Y. *et al.*, 2013; Moreau, Y. *et al.*, 2012).

Here, we consider the set of known breast cancer genes (129 genes) in Table A.2 as seed genes, and subsequently used this set to prioritize certain genes throughout this analysis. We applied a straightforward gene prioritization strategy, to increase our chance of regaining interactions between known and probably yet unknown genes for breast cancer from our dataset. Thus, for each module, if there exists a path of distance equal to the average distance, i.e. the absolute value of 3.9, between a seed gene and the central gene of the module, then the seed gene from the set of known breast cancer genes is added to the module. We observed that the modules contained too many genes with too much redundancy among the modules, making the analysis less manageable. Therefore, to reduce redundancy among our modules during the

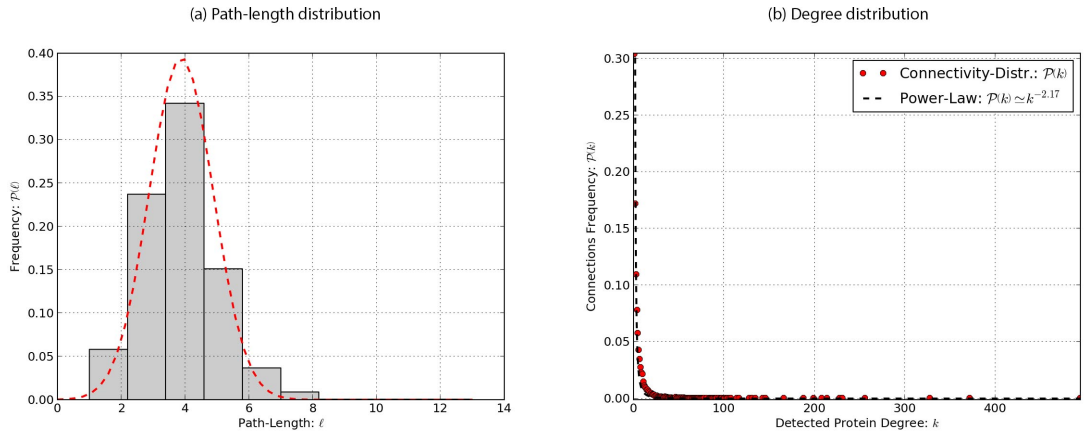

Figure 4.1: Topological analysis of the CGEMS breast cancer LD-weighted PPI network. **(a)** Distribution of shortest path lengths between reachable pair-wise protein functional interactions. **(b)** Connectivity distribution of detected k functional links per protein, plotted as a function of frequency $\mathcal{P}(k)$.

prioritization process, if a seed gene has already been attached to a module, it will not be attached to another module with the same interactions. This did not affect the results as the pathway enrichment results were almost the same from these two strategies. Finally, because over 90% of the pathways from BioCarta in our database had < 30 genes, while other pathways from databases such as KEGG and PID were generally larger and more variable in size (mean gene count $= 44$ and 36, standard deviation gene count $= 24$ and 16 respectively), we decided to only retain pathways with size ranges from 10 to 100. This was to avoid bias due to large pathways, having greater chances of being represented just because of containing a much larger number of genes.

A total of 596 sub-networks (modules) were generated from ancGWAS, and the significance of each module was assessed using the Stouffer-Liptak method. From the 20 top sub-networks obtained from ancGWAS, we performed enrichment analysis on each sub-network using our pathway database of $1,047$ annotated pathways (see Subsection 2.1.7). Table 4.4 presents the top 20 module hubs from ancGWAS analysis, ranked by their Stouffer-Liptak p-values, as well as the pathway with the highest z-score for each module from our enrichment analysis. Interestingly, 11 pathways were significantly overlapping with our top 20 modules with higher amount of overlap, many of them containing at least 1 known breast cancer disease gene. Of these, *Proteoglycan syndecan-1-mediated signaling events* from the PID database was the most frequent, representing almost 35% of pathways mapped to the top 20 modules for breast cancer in this analysis. This pathway was also identified as the most enriched in a previous pathway-based analysis study of breast cancer by Menashe in (Menashe, I. *et al.*, 2010) using the same dataset as this study, but using an enrichment score reflecting the overrepresentation of gene-based association signals in each pathway using a weighted Kolmogorov-Smirnov procedure.

Table 4.4: Top 20 sub-networks, and related pathway enrichment results from ancGWAS. For each pathway, the z-score (ZP) is reported representing the level of its similarity with the related module, which is also a function of the size of the module. Different scores of overlaps between the module, the pathway and the set of known genes are also supplied. The list of module genes is provided in Table A.4 in supplementary materials.

Hub	P	$AdjP$	**MS**	**Pathway**	**ZP**	**MOG**	**MKG**	**PKG**
MAP3K12	2.69×10^{-07}	0.19185	17	Beta1 integrin cell surface interactions	1.318	4	2	1
CDKN1A	1.01×10^{-05}	0.14154	22	Proteoglycan syndecan-mediated signaling events	2.7858	6	2	1
UNC93B1	3.70×10^{-05}	0.12664	13	Metabolism of proteins	35.159	2	1	0
EIF2B1	3.97×10^{-05}	0.17215	19	Beta1 integrin cell surface interactions	2.0834	5	2	1
SMC3	7.61×10^{-05}	0.11893	26	Cell Cycle, Mitotic	20.865	5	1	1
LNX2	0.00030	0.11164	27	Proteoglycan syndecan-mediated signaling events	5.3384	11	1	1
ITGA5	0.00064	0.13200	11	Proteoglycan syndecan-mediated signaling events	1.2439	3	1	1
ATF7IP	0.00085	0.13739	14	Beta1 integrin cell surface interactions	3.7639	3	1	1
GRIN2B	0.00154	0.13751	20	Glypican	6.0553	9	2	2
RTN4	0.00167	0.09724	14	Proteoglycan syndecan-mediated signaling events	5.3186	6	3	1
PRKCG	0.00210	0.11598	19	ErbB receptor signaling network	2.7864	4	1	1
MAPKAP1	0.00390	0.10279	20	Integrin family cell surface interactions	5.7685	9	2	1
PACSIN3	0.00513	0.11618	20	Beta1 integrin cell surface interactions	6.5632	8	1	1
LEF1	0.00560	0.07560	21	Proteoglycan syndecan-mediated signaling events	4.6443	9	4	1
ACVR1	0.00786	0.11361	21	CDC42 signaling events	10.462	5	1	1
PCBD2	0.01319	0.11920	11	Gene Expression	16.282	2	1	0
WAS	0.01377	0.09055	13	Proteoglycan syndecan-mediated signaling events	4.6116	5	1	1
BSG	0.01920	0.08244	16	Proteoglycan syndecan-mediated signaling events	1.7079	3	2	1
ARL4D	0.01972	0.11676	12	Regulation of Androgen receptor activity	21.888	1	1	0
FBXW11	0.02184	0.07183	21	Signaling events mediated by *VEGFR1* and *VEGFR2*	1.9113	4	2	2

Abbreviation. P: P-value; $AdjP$: Adjusted p-value; MS: Module size; ZP: Pathway z-score; MOG: Number of genes overlapping between the module and the pathway; MKG: Number of genes overlapping between the module and the set of known breast cancer genes; PKG: Number of genes overlapping between the pathway and the set of known breast cancer genes;

The pathway is believed to be involved in breast cancer development (Menashe, I. *et al.*, 2010). Other interesting pathways include the *ErbB receptor signaling network*, also known as the *HER Signaling Pathway*, which is one of the most important breast cancer pathways used for targeted therapies for breast cancer (*The HER2 Pathway in Breast Cancer.* „ National Cancer

Institute, 2014); the *Glypican* pathway, which has also shown significant overlap with one of our modules, with 2 known breast cancer genes, and has also been shown to be involved in cancer (Filmus, J., 2001), and the *Regulation of Androgen receptor activity* pathway. While having moderate p-values with less overlap with the ancGWAS module and known breast cancer disease genes, the latter pathway contains genes known to play a role in normal breast physiology, and is becoming increasingly recognized as an important contributor towards breast carcinogenesis (Garay, J. P. *et al.*, 2012).

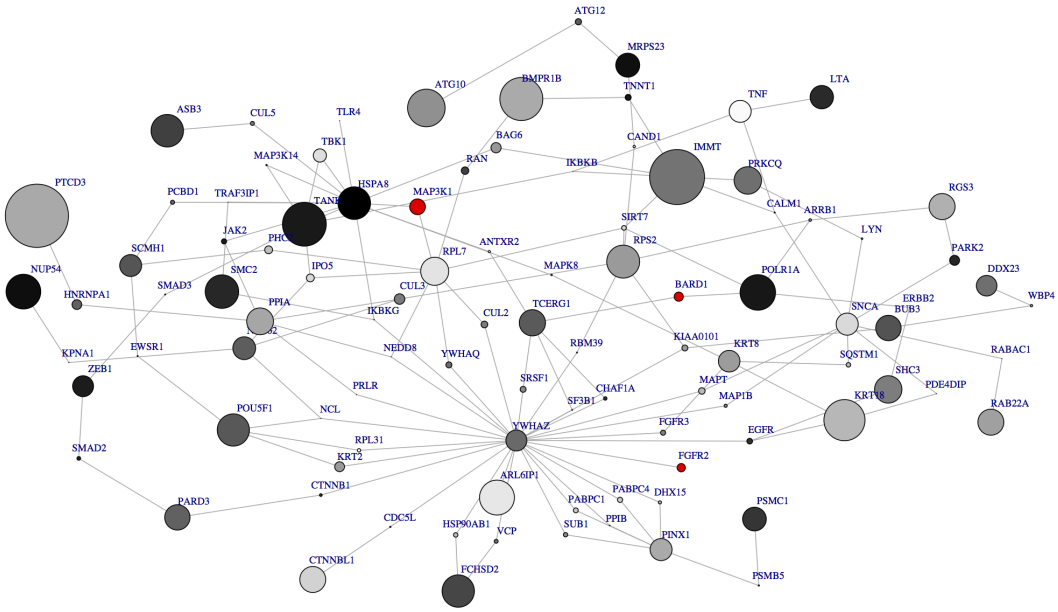

Figure 4.2: Central network of 100 genes for breast cancer dataset. The size of a node denotes its significance with size increasing with significance. In red are known breast cancer disease genes.

Moreover, we observed that the 20 top sub-networks obtained from ancGWAS also overlap each other, and most of the hubs are connected directly or indirectly to each other. In light of this considerable overlap between these modules in our study, we can hypothesize that common biological modules (pathways) may underlie the enrichment signals in multiple biological pathways. To explore this hypothesis, we searched for the most important and central sub-network (Figure 4.2) within the network formed from these 20 sub-networks. For ease of presentation, we reduced the resulting sub-network by excluding those genes with less than three edges (neighbours) and applying a recursive approach. Figure 4.2 shows the most important sub-network related to breast cancer, it contains previously associated breast cancer genes, particularly the *FGFR2* gene ($p = 0.037$), which was the most significant gene from the CGEMS study. Though *FGFR2* has a moderate association signal, it is interacting with an important gene (with respect to its connectedness) in this network, the antiapoptotic YWHAZ gene, which has 27 partners, and is holding together many other important genes in this

network, as well as connecting several components of this network. This gene (*YWHAZ*) has not yet been associated with breast cancer, even after combining p-values from multiple single SNPs ($p = 0.00024$), however it is a key mediator of different pathways illuminated in this study and principally in the central sub-network for breast cancer here. A recent study has demonstrated its involvement in *de novo* chemoresistance to anthracyclines and its permissibility for metastatic recurrence (Li, Y. *et al.*, 2010). Therefore, it may modulate breast cancer susceptibility through different biological mechanisms.

Besides *FGFR2*, two additional known breast cancer disease genes are part of this central sub-network. The *MAP3K1* gene (also known as *MEKK1*), which encodes the *MAPK* kinase protein that phosphorylates and activates the *MAPK* kinase (*MAPK2*), that, in turn, phosphorylates the *MAPK/ERK* to produce downstream signaling effects on a variety of cancer genes. It is principally a member of the *MAPK* pathway, strongly associated with *HER* receptor activity, the mutations of which have been associated with HER2+ breast tumors (Rebbeck, T. R. *et al.*, 2009). The *BARD1* gene forms heterodimers with the breast cancer gene *BRCA1*, and it is speculated that BARD1 mutations might affect the function of BRCA1, contributing to breast carcinogenesis (Ishitobi, M. *et al.*, 2003). Looking at some non breast cancer disease genes (i.e. genes not yet confirmed to be associated with breast cancer), the *RPL7* gene, which is part of the *Metabolism of proteins* and *Gene expression pathways* (A.4 and supplementary Table A.1), interacts with gene *HSPA8*, which plays a role in cell growth. It is believed to be involved in metastasis of breast cancer, yet interacts with gene *TLR4*, a member of the Toll-like receptor signaling pathway (*TLRs*) involved in the production of pro-inflammatory cytokines with activation of $NF_{\kappa}B$ and MAPK. They are associated with the induction of *IFN*-beta and *IFN*-inducible genes, and maturation of dendritic cells. TLRs are expressed on innate immune cells, such as macrophages and dendritic cells (Akira, S. *et al.*, 2004). The *RPL7* gene, or any other gene in this sub-network, may have a role in the breast carcinogenesis, but was not detected through the single-marker-based approach of GWAS using this dataset. This illustrates the benefit of incorporating both the association signal from a standard GWAS and the human PPI network for testing the combined effects of SNPs and searching for significantly enriched sub-networks for complex diseases. This had led to new hypothesis of breast cancer-associated genes, but these would need to be validated.

Finally, we performed a pathway enrichment analysis of this network, and reported the pathway with the highest score of overlap with our central sub-network. This enrichment resulted in the *Integrin family cell surface interactions* pathway, which was also enriched in our top 20 modules in Table 4.4. Integrins are a family of transmembrane glycoprotein receptors that mediate cell-matrix and cell-cell interactions, consisting of an α and $\alpha\beta$ subunits, and they are known for having a pivotal role in cellular behavior, as well as in many pathological conditions such as inflammation and tumor progression. Several studies have demonstrated the

association between the regulation of integrin expression and cancer, as well as breast cancer. Links between the integrins and other important pathways contributing to the development of breast tumors have also been investigated such as the hepatocyte growth factor/scatter factor (HGF/SF) whereby integrins $\beta 1$, $\beta 3$, $\beta 4$ and $\beta 5$ may be affected by HGF-mediated regulation of integrin avidity, and insulin-like growth factor I ($IGF\text{-}I$), which stimulates an increase in the activity of integrin $\alpha V\beta 5$. The epidermal growth factor and its receptor ($EGFR$) have also been reported to mediate up-regulation of $\beta 1$ integrin function and breast cancer progression (Koistinen, P. *et al.*, 2000).

Our results suggest that genetic alterations associated with the top 20 pathways and the central pathway for our study may contribute to breast cancer susceptibility. While, some of these pathways are already known and targeted for breast cancer therapies, others may be new, and harbouring genes that participate actively in biological processes involved in breast cancer. These pathways can, in addition, reveal novel genes to target as potential candidates for association studies aiming at understanding breast carcinogenesis. Ultimately, further studies would definitively be needed to confirm these results, and explore further the genetic variations underlying the association of these pathways with breast cancer.

Most targeted therapies that are currently being developed for breast cancer are used and tested in combination with standard therapies. However, since cancer cells are often subject to multiple signaling pathways, targeting multiple pathways for a combination of targeted therapies might reduce the development of drug-resistant tumor cells, and help in designing optimal dosages and schedules for combination therapies (National Cancer Institute, 2014).

CHAPTER 5

Conclusion

As stated in the introduction to this thesis, the primary purpose of the study was to explore post Genome-wide Association Study approaches for genetic association studies with respect to the combined effect of interacting loci, and the implementation of a new network-based method to identify sub-networks of genes associated with complex disease in admixed heterogeneous population, as well as in homogeneous populations. As the overview in the theoretical part of this thesis showed, the need for complementary methods for current genome-wide association study methods is imperative. On the other hand, although our knowledge on the human interactome is far from complete, integrating network topological features with GWAS data can provide experimentally verifiable insights into the understanding of complex traits.

In the first chapter of this thesis, I introduced the background and literature relevant to this project. We discussed basic concepts on genetic association studies and genome-wide association study, focusing on the case-control design in population association studies. The limitations of current methods for genome-wide association study motivated us to look at alternative approaches that can complement the single-marker-based approach of GWAS, further developed in chapter two. We also reviewed current methods for pathway-based analysis of GWAS data, focusing on some technical differences between these pathway-based approaches, as well some common challenges observed in these methods. I have also discussed some functional information that is widely used in pathway-based methods for GWA study, the comprehension of which was vital for the design and implementation of the new method, ancGWAS, in chapter 2. The main aim of chapter three was to propose a new method to identify sub-networks that combined effects of multiple interacting loci based of the topological structure of the gene-gene network that results from GWA study data, also incorporating the ancestral information in the case of an admixed population. In the fourth chapter, we presented ancGWAS and the implemented algorithm for a step-wise network-based analysis of GWA study data. We also emphasized testing for genes and pathways specific for particular subpopulations by integrating ancestral information into the analysis. Besides, this method has the advantage of

using the structure of the network generated using known human protein-protein interactions from the genes from GWAS, and uses the correlation between these genes to investigate genes playing important structural roles (which can also represent biological roles), to break down the network into sub-networks, before scoring these sub-networks using GWA study signals. Thus, ancGWAS sub-networks can not only discover novel disease pathways, but also reveal disease genes that are currently not part of any pathways. Through simulations in chapter four, we assessed how well ancGWAS performs in identifying disease associated pathways, and demonstrated that ancGWAS performs better than an existing network-based method, dmGWAS, and holds promise for approximating disease associated pathways. We also tested for differences in ancestry at gene and pathway levels, to investigate whether deviations from the expected ancestry proportions in genes and generated sub-networks exist. Chapter five provides an application of the proposed new method on a real data of case-control breast cancer study. The majority of sub-networks identified by ancGWAS were highly enriched in known breast cancer genes, and overlapping in genes with well known breast cancer pathways from public pathway annotation databases. This suggests that the presence of many of the genes in these sub-networks, which are not known breast cancer genes may be because of the incompleteness of pathway databases and our incomplete knowledge of disease associated genes, since they are connected with short paths to known breast cancer genes, and thus may be involved in the disease pathogenesis.

Although ancGWAS has been demonstrated using a GWAS of breast cancer, it can be used to interpret GWA studies of other traits and organisms too. By identifying disease associated modules, the ancGWAS approach provides insights into the involvement in complex phenotypes with multiple susceptibility loci with small effects. In this way, ancGWAS as well as other network-based approaches may be of crucial use in gaining a thorough understanding of biological function of crucial genetic components in order to dissect complex diseases in the coming era of systems medicine.

Despite many challenges of pathways-based approaches for analysis of GWAS data, which need to be addressed as discussed in chapter 1, there are also many opportunities ahead. For instance, the integration of additional types of data to SNP genotypes, including copy number variants, gene expression, epigenetic modifications and somatic mutations, among other genomic data types, together into the same pathway analysis may be more powerful in revealing novel biological insights. In addition to improvements in the searching algorithm, future work will include integrating other types of genomic data, including the Gene Ontology (GO), and integrating data sets from multiple GWA studies, as is done in meta-analysis, into one pathway-based analysis tool to increase the power to reveal biological insights.

Bibliography

1. Akira, S. *et al.*, "Toll-like receptor signalling", *Nat Rev Immunol*, vol. 4, no. 7, pp. 499–511, 2004.

2. A.L., B. *et al.*, "Emergence of scaling in random networks.", *Science*, vol. 286, pp. 509–512, 1999.

3. Albert, R. *et al.*, "Statistical mechanics of complex networks", *Rev. Mod. Phys.* Vol. 74, no. 1, pp. 47–97, 2002.

4. Anderson, C. A. *et al.*, "Data quality control in genetic case-control association studies", *Nat Protoc*, vol. 5, no. 9, pp. 1564–73, 2010.

5. Anderson, C. A. *et al.*, "Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47", *Nat Genet*, vol. 43, no. 3, pp. 246–52, 2011.

6. Andrea, S. F., *Applied Statistical Genetics with R*, Use R, Springer New York, 2009, pp. 1–27.

7. Astle, W. *et al.*, "Population Structure and Cryptic Relatedness in Genetic Association Studies", *Statistical Science*, vol. 24, no. 4, pp. 451–471, 2009.

8. Aulchenko, Y. S. *et al.*, "GenABEL: an R library for genome-wide association analysis", *Bioinformatics*, vol. 23, no. 10, pp. 1294–6, 2007.

9. Bader, G. D. *et al.*, "Pathguid:: a pathway resource list", *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D504–6, 2006.

10. Bao, S. Y. *et al.*, "Prioritizing genes responsible for host resistance to influenza using network approaches", *Bmc Genom*, vol. 14 2013.

11. Baran, Y. *et al.*, "Fast and accurate inference of local ancestry in Latino populations", *Bioinformatics*, vol. 28, no. 10, pp. 1359–67, 2012.

12. Baranzini, S. E. *et al.*, "Pathway and network-based analysis of genome-wide association studies in multiple sclerosis", *Hum Mol Genet*, vol. 18, no. 11, pp. 2078–90, 2009.

13. Barrett, J. C. *et al.*, "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease", *Nat Genet*, vol. 40, no. 8, pp. 955–62, 2008.

14. Barrett, J. C. *et al.*, "Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes", *Nat Genet*, vol. 41, no. 6, pp. 703–707, 2009.

15. Baselga, J., "Targeting the phosphoinositide-3 (PI3) kinase pathway in breast cancer", *Oncologist*, vol. 16 Suppl 1, pp. 12–9, 2011.

16. Benjamini, Y. *et al.*, "Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing", *J Roy Stat Soc Series B-Meth*, vol. 57, no. 1, pp. 289–300, 1995.

17. Berger, S. I. *et al.*, "Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases", *BMC Bioinformatics*, vol. 8 2007.

18. Bigby, M., "Odds Ratios and Relative Risks", *Arch Dermatol*, vol. 136, no. 6, p. 770, 2000.

19. Bjorn, H. J. *et al.*, *Analysis of Biological Networks (Wiley Series in Bioinformatics)*, New York, NY, USA: Wiley-Interscience, 2008.

20. Briollais, L. *et al.*, "Methodological issues in detecting gene-gene interactions in breast cancer susceptibility: a population-based study in Ontario", *BMC Med*, vol. 5, p. 22, 2007.

21. Burton, P. R. *et al.*, "Key concepts in genetic epidemiology", *Lancet*, vol. 366, no. 9489, pp. 941–51, 2005.

22. Bush, W. S. *et al.*, "Chapter 11: Genome-wide association studies", *PLoS Comput Biol*, vol. 8, no. 12, e1002822, 2012.

23. Caldwell, R. *et al.*, *Genetic Drift: Bottlenecks and Founder Effects*, http://evolution.berkeley.edu/evosite/evo101/IIID3Bottlenecks.shtml, [Online; accessed March2014], 2014.

24. Cantor, R. M. *et al.*, "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application", *Am J Hum Genet*, vol. 86, no. 1, pp. 6–22, 2010.

25. Casci, T., "Association studies: The best of the rest", *Nature Reviews Genetics*, vol. 8, no. 12, pp. 907–907, 2007.

26. Chang, E. *et al.*, *Quiz Answers: Heritability Demystified.* http://blog.23andme.com/23andmeand-you/genetics101/quizanswersheritabilitydemystified/, [Online; accessed March-2014], 2014.

27. Chang, J. T. *et al.*, "A genomic strategy to elucidate modules of oncogenic pathway signaling networks", *Mol Cell*, vol. 34, no. 1, pp. 104–14, 2009.

28. Chatr-aryamontri, A. *et al.*, "MINT: the Molecular INTeraction database", *Nucleic Acids Res*, vol. 35, no. Database issue, pp. D572–4, 2007.

29. Chen, X. *et al.*, "Pathway-based analysis for genome-wide association studies using supervised principal components", *Genet Epidemiol*, vol. 34, no. 7, pp. 716–24, 2010.

30. Chen, Y. *et al.*, "Variations in DNA elucidate molecular networks that cause disease", *Nature*, vol. 452, no. 7186, pp. 429–35, 2008.

31. Chimusa, E. R. *et al.*, "Genome-wide association study of ancestry-specific TB risk in the South African Coloured population", *Hum Mol Genet*, vol. 1, no. 1 2013.

32. Chimusa, E. R. *et al.*, "Genome-wide association study of ancestry-specific TB risk in the South African Coloured population", *Hum Mol Genet*, vol. 23, no. 3, pp. 796–809, 2014.

33. Cho, M. *et al.*, "STAT3 and NF-kB Signal Pathway Is Required for IL-23-Mediated IL-17 Production in Spontaneous Arthritis Animal Model IL-1 Receptor Antagonist-Deficient Mice", *PNAS*, vol. 103, no. 5, pp. 1446–1451, 2006.

34. Choi, S. C., "Tests of Equality of Dependent Correlation Coefficients", *Biometrika Trust*, vol. 64, no. 3, pp. 645–647, 1977.

35. Chuang, H. Y. *et al.*, "Network-based classification of breast cancer metastasis", *Mol Syst Biol*, vol. 3, p. 140, 2007.

36. Collins, A. *et al.*, "The genetics of breast cancer: risk factors for disease", *Appl Clin Genet*, vol. 4, no. 1, pp. 11–9, 2011.

37. Cordell, H. J. *et al.*, "Genetic association studies", *Lancet*, vol. 366, no. 9491, pp. 1121–1131, 2005.

38. Cowley, M. J. *et al.*, "PINA v2.0: mining interactome modules", *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D862–5, 2012.

39. Dang, C. *et al.*, *The HER2 Pathway in Breast Cancer.* http://am.asco.org/her2-pathway-breast-cancer, [Online; accessed March2014].

40. de Wit, E. *et al.*, "Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape", *Hum Genet*, vol. 128, no. 2, pp. 145–53, 2010.

41. de Wit, E. *et al.*, "Gene-gene interaction between tuberculosis candidate genes in a South African population", *Mamm Genome*, vol. 22, no. 1-2, pp. 100–10, 2011.

42. Devlin, B. *et al.*, "A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping", *Genomics*, vol. 29, no. 2, pp. 311–322, 1995.

43. Devlin, B. *et al.*, "Genomic Control for Association Studies", *Biometrics*, vol. 55, no. 4, pp. 997–1004, 1999.

44. Dries, D. L., "Genetic ancestry, population admixture, and the genetic epidemiology of complex disease", *Circ Cardiovasc Genet*, vol. 2, no. 6, pp. 540–3, 2009.

45. Edwards, A. O. *et al.*, "Complement factor H polymorphism and age-related macular degeneration", *Science*, vol. 308, no. 5720, pp. 421–4, 2005.

46. Fehringer, G. *et al.*, "Comparison of pathway analysis approaches using lung cancer GWAS data sets", *PLoS One*, vol. 7, no. 2, e31816, 2012.

47. Feng, Z. *et al.*, "PATHSIMU: A Flexible Simulating Tool for Pathway-based Genome-wide Association Studies", *Open Acc Sc Rep*, vol. 1, no. 1 2012.

48. Filmus, J., "Glypicans in growth control and cancer", *Glycobiology*, vol. 11, no. 3, 19R–23R, 2001.

49. Fisher, R., "Statistical Methods for Research Workers", *Am Math*, vol. 37, no. 10, pp. 547–550, 1958.

50. Folks, J., "Combination of independent tests", *in P. Krishnaiah, ed.* Vol. 4, pp. 113–121, 1984.

51. Franke, A. *et al.*, "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci", *Nat Genet*, vol. 42, no. 12, pp. 1118–25, 2010.

52. Frazer, K. A. *et al.*, "Human genetic variation and its contribution to complex traits", *Nat Rev Genet*, vol. 10, no. 4, pp. 241–51, 2009.

53. Freidlin, B. *et al.*, "Trend tests for case-control studies of genetic markers: power, sample size and robustness", *Hum Hered*, vol. 53, no. 3, pp. 146–52, 2002.

54. Freimer, N. B. *et al.*, "Human genetics: variants in common diseases", *Nature*, vol. 445, no. 7130, pp. 828–30, 2007.

55. Fridley, B. L. *et al.*, "Gene set analysis of SNP data: benefits, challenges, and future directions", *Eur J Hum Genet*, vol. 19, no. 8, pp. 837–43, 2011.

56. Gao, X. *et al.*, "Transition Dependency: A Gene-Gene Interaction Measure for Times Series Microarray Data", *EURASIP J Bioinform Syst Biol*, vol. 2009, no. 1, p. 535869, 2009.

57. Garay, J. P. *et al.*, "Androgen receptor as a targeted therapy for breast cancer", *Am J Cancer Res*, vol. 2, no. 4, pp. 434–445, 2012.

58. Goh, K. I. *et al.*, "The human disease network", *Proc Natl Acad Sci U S A*, vol. 104, no. 21, pp. 8685–90, 2007.

59. Goldstein, D. B. *et al.*, "Population genomics: linkage disequilibrium holds the key", *Curr Biol*, vol. 11, no. 14, R576–9, 2001.

60. Gui, H. *et al.*, "Comparisons of seven algorithms for pathway analysis using the WTCCC Crohnś Disease dataset", *BMC Res Notes*, vol. 4, p. 386, 2011.

61. Guo, Y. F. *et al.*, "A new permutation strategy of pathway-based approach for genome-wide association study", *BMC Bioinform*, vol. 10, no. 1, p. 429, 2009.

62. Han, B. *et al.*, "Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies", *Am J Hum Genet*, vol. 88, no. 5, pp. 586–598, 2011.

63. Hanahan, D. *et al.*, "The hallmarks of cancer", *Cell*, vol. 100, pp. 57–70, 2000.

64. Hancock, D. B. *et al.*, "Assessment of genotype imputation performance using 1000 Genomes in African American studies", *PLoS One*, vol. 7, no. 11, e50610, 2012.

65. Hardy, J. *et al.*, "Genomewide association studies and human disease", *N Engl J Med*, vol. 360, no. 17, pp. 1759–68, 2009.

66. Hart, G. T. *et al.*, "How complete are current yeast and human protein-interaction networks?", *Genome Biol*, vol. 7, no. 11, p. 120, 2006.

67. Hedges, L. *et al.*, *Statistical Methods for Meta-Analysis*, London: Academic Press, 1985.

68. Henn, B. M. *et al.*, "Genomic ancestry of North Africans supports back-to-Africa migrations", *PLoS Genet*, vol. 8, no. 1, e1002397, 2012.

69. Hermjakob, H. *et al.*, "IntAct: an open source molecular interaction database", *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D452–5, 2004.

70. Hess, A. *et al.*, "Fisher's Combined P-value for Detecting Differentially Expressed Genes using Affymetrix Expression Arrays", vol. 8, no. 96 2007.

71. Hindorff, L. A. *et al.*, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits", *Proc Natl Acad Sci U S A*, vol. 106, no. 23, pp. 9362–7, 2009.

72. Hindorff, L. A. *et al.*, *A catalog of published genome-wide association studies*, http://www.genome.gov/gwastudies, [Online; accessed March2014], 2013.

73. Hindorff, L. A. *et al.*, *A Catalog of Published Genome-Wide Association Studies.* www.genome.gov/gwastudies, [Online; accessed March2014], 2013.

74. Hoggart, C. J. *et al.*, "Control of confounding of genetic associations in stratified populations", *Am J Hum Genet*, vol. 72, no. 6, pp. 1492–1504, 2003.

75. Hoh, J. *et al.*, "Trimming, weighting, and grouping SNPs in human case-control association studies", *Genome Res*, vol. 11, no. 12, pp. 2115–9, 2001.

76. Holmans, P. *et al.*, "Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder", *Am J Hum Genet*, vol. 85, no. 1, pp. 13–24, 2009.

77. Howie, B. N. *et al.*, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies", *PLoS Genet*, vol. 5, no. 6, e1000529, 2009.

78. Hu, P. *et al.*, "Pathway-based joint effects analysis of rare genetic variants using Genetic Analysis Workshop 17 exon sequence data", *BMC Proc*, vol. 5 Suppl 9, no. 1, S45, 2011.

79. Hunter, D. J. *et al.*, "A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer", *Nat Genet*, vol. 39, no. 7, pp. 870–4, 2007.

80. Ideker, T. *et al.*, "Discovering regulatory and signalling circuits in molecular interaction networks", *Bioinformatics*, vol. 18 Suppl 1, S233–40, 2002.

81. Iles, M. M., "What can genome-wide association studies tell us about the genetics of common disease?", *PLoS Genet*, vol. 4, no. 2, e33, 2008.

82. International HapMap Consortium, "Integrating common and rare genetic variation in diverse human populations", *Nature*, vol. 467, no. 7311, pp. 52–8, 2010.

83. International Multiple Sclerosis Genetics Consortium, "Network-Based Multiple Sclerosis Pathway Analysis with GWAS Data from 15,000 Cases and 30,000 Controls", *Am J Hum Genet*, vol. 1, no. 1 2013.

84. International Parkinson Disease Genomics Consortium, "Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies", *Lancet*, vol. 377, no. 9766, pp. 641–649, 2011.

85. Ishitobi, M. *et al.*, "Mutational analysis of BARD1 in familial breast cancer patients in Japan", *Cancer Lett*, vol. 200, no. 1, pp. 1–7, 2003.

86. Jensen, M. K. *et al.*, "Protein interaction-based genome-wide analysis of incident coronary heart disease", *Circ Cardiovasc Genet*, vol. 4, no. 5, pp. 549–56, 2011.

87. Jia, P. *et al.*, "dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks", *Bioinformatics*, vol. 27, no. 1, pp. 95–102, 2011.

88. Jia, P. *et al.*, "A bias-reducing pathway enrichment analysis of genome-wide association data confirmed association of the MHC region with schizophrenia", *J Med Genet*, vol. 49, no. 2, pp. 96–103, 2012.

89. John, F. Y. B., "Q&A: Promise and pitfalls of genome-wide association studies", *BMC Biology*, vol. 8, no. 1, p. 41, 2010.

90. Kang, H. M. *et al.*, "Variance component model to account for sample structure in genome-wide association studies", *Nat Genet*, vol. 42, no. 4, pp. 348–54, 2010.

91. Karp, P. D. *et al.*, "EcoCyc: Encyclopedia of Escherichia coli genes and metabolism", *Nucleic Acids Res*, vol. 26, no. 1, pp. 50–3, 1998.

92. Keshava Prasad, T. S. *et al.*, "Human Protein Reference Database–2009 update", *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D767–72, 2009.

93. Khoury, M. J. *et al.*, *Fundamentals of Genetic Epidemiology*, Oxford University Press, 1993, p. 400.

94. Kittles, R. A. *et al.*, "Race, ancestry, and genes: implications for defining disease risk", *Annu Rev Genomics Hum Genet*, vol. 4, no. 1, pp. 33–67, 2003.

95. Klein, C. *et al.*, "The Promise and Limitations of Genome-wide Association Studies", *JAMA*, vol. 308, no. 18, pp. 1867–1868, 2012.

96. Klein, R. J. *et al.*, "Complement factor H polymorphism in age-related macular degeneration", *Science*, vol. 308, no. 5720, pp. 385–9, 2005.

97. Knowler, W. C. *et al.*, "Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture.", *Am J Hum Genet.* Vol. 1, no. 1 1988.

98. Knudson, A. G., Jr., "Mutation and cancer: statistical study of retinoblastoma", *Proc Natl Acad Sci U S A*, vol. 68, no. 4, pp. 820–3, 1971.

99. Koistinen, P. *et al.*, *In: Madame Curie Bioscience Database*, http://www.ncbi.nlm.nih.gov/books/NBK6070/, [Online; accessed March2014], 2000.

100. Kortylewski, M. *et al.*, "Regulation of the IL-23 and IL-12 balance by Stat3 signaling in the tumor microenvironment", *Cancer Cell*, vol. 15, no. 2, pp. 114–123, 2009.

101. Kost, J. T. *et al.*, "Combining dependent *p*-values", *Elsevier Sc*, vol. 60, no. 2, pp. 183–190, 2002.

102. Kraft, P. *et al.*, "Complex diseases, complex genes: keeping pathways on the right track", *Epidemiology*, vol. 20, no. 4, pp. 508–11, 2009.

103. Krauthammer, M. *et al.*, "Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease", *Proc Natl Acad Sci U S A*, vol. 101, no. 42, pp. 15148–53, 2004.

104. Lehne, B. *et al.*, "Protein-protein interaction databases: keeping up with growing interactomes", *Hum Genomics*, vol. 3, no. 3, pp. 291–7, 2009.

105. Lehne, B. *et al.*, "From SNPs to genes: disease association at the gene level", *PLoS One*, vol. 6, no. 6, e20133, 2011.

106. Leivonen, M. *et al.*, "Prognostic value of syndecan-1 expression in breast cancer", *Oncology*, vol. 67, no. 1, pp. 11–8, 2004.

107. Lewis, C. M., "Genetic association studies: Design, analysis and interpretation", *Brief in Bioinform*, vol. 3, no. 2, pp. 146–153, 2002.

108. Lewis, C. M. *et al.*, "Introduction to genetic association studies", *Cold Spring Harb Protoc*, vol. 2012, no. 3, pp. 297–306, 2012.

109. Li, H. *et al.*, "Inference of human population history from individual whole-genome sequences", *Nature*, vol. 475, 493âĂŞ496, 2011.

110. Li, J. Z. *et al.*, "Worldwide human relationships inferred from genome-wide patterns of variation", *Science*, vol. 319, no. 5866, pp. 1100–4, 2008.

111. Li, W., "Three lectures on case-control genetic association analysis", *Brief Bioinform*, vol. 9, no. 1, pp. 1–13, 2008.

112. Li, Y. *et al.*, "Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer", *Nat Med*, vol. 16, no. 2, pp. 214–8, 2010.

113. Liang, H. *et al.*, "Gene essentiality, gene duplicability and protein connectivity in human and mouse", *Trends Genet*, vol. 23, no. 8, pp. 375–8, 2007.

114. Liptak, T., "On the Combination of Independent Tests", *Magyar*, vol. 3, pp. 171–197, 1958.

115. Loughin, T. M., "A systematic comparison of methods for combining p-values from independent tests", *Computational Statistics & Data Analysis*, vol. 47, no. 3, pp. 467–485, 2004.

116. Lunetta, K. L., "Genetic association studies", *Circulation*, vol. 118, no. 1, pp. 96–101, 2008.

117. Ma'ayan, A., "Introduction to network analysis in systems biology", *Sci Signal*, vol. 4, no. 190, tr5, 2011.

118. Ma'ayan, A. *et al.*, "Formation of regulatory patterns during signal propagation in a Mammalian cellular network", *Science*, vol. 309, no. 5737, pp. 1078–83, 2005.

119. H. Maciejewski, "Competitive and self-contained gene set analysis methods applied for class prediction", *FedCSIS, 2011*, pp. 55–61.

120. Manolio, T. A. *et al.*, "Finding the missing heritability of complex diseases", *Nature*, vol. 461, no. 7265, pp. 747–53, 2009.

121. Martinez, A. *et al.*, "Epistatic interaction between FCRL3 and NFkappaB1 genes in Spanish patients with rheumatoid arthritis", *Ann Rheum Dis*, vol. 65, no. 9, pp. 1188–91, 2006.

122. Mathew, C. G., "New links to the pathogenesis of Crohn disease provided by genome-wide association scans", *Nat Rev Genet*, vol. 9, no. 1, pp. 9–14, 2008.

123. Mazandu, G. K. *et al.*, "Generation and Analysis of Large-Scale Data-Driven Mycobacterium tuberculosis Functional Networks for Drug Target Identification", *Adv Bioinformatics*, vol. 2011, p. 801478, 2011.

124. McCarthy, M. I., "Genomics, type 2 diabetes, and obesity", *N Engl J Med*, vol. 363, no. 24, pp. 2339–50, 2010.

125. McCarthy, M. I. *et al.*, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges", *Nat Rev Genet*, vol. 9, no. 5, pp. 356–69, 2008.

126. Menashe, I. *et al.*, "Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade", *Cancer Res*, vol. 70, no. 11, pp. 4453–9, 2010.

127. Moreau, Y. *et al.*, "Computational tools for prioritizing candidate genes: boosting disease gene discovery", *Nat Rev Genet*, vol. 13, no. 8, pp. 523–36, 2012.

128. Morton, N. E., *Outline of genetic epidemiology*, London: S Karger Ag, 1982, p. 252.

129. National Cancer Institute, *Targeted Therapies for Breast Cancer Tutorial.* http://www.cancer.gov/cancertopics/understandingcancer/targetedtherapies/breastcancer_htmlcourse, [Online; accessed March2014], 2014.

130. National Institute of Health, *Genetics Home Reference: Breast cancer.* http://ghr.nlm.nih.gov/condition/breast-cancer, [Online; accessed March-2014], 2014.

131. Nature Genetics, "On beyond GWAS", *Nat Genet*, vol. 42, no. 7, p. 551, 2010.

132. Newman, M. E., "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality", *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 64, no. 1 Pt 2, p. 016132, 2001.

133. Ostrer, H., "A genetic profile of contemporary Jewish populations", *Nat Rev Genet*, vol. 2, no. 11, pp. 891–8, 2001.

134. Pavlopoulos, G. A. *et al.*, "Using graph theory to analyze biological networks", *BioData Min*, vol. 4, p. 10, 2011.

135. Pearson, T. A. *et al.*, "How to interpret a genome-wide association study", *JAMA*, vol. 299, no. 11, pp. 1335–44, 2008.

136. Peng, G. *et al.*, "Gene and pathway-based second-wave analysis of genome-wide association studies", *Eur J Hum Genet*, vol. 18, no. 1, pp. 111–7, 2010.

137. Penny, P. D. *et al.*, *Variational bayes for 1-dimensional mixture models*, Report, University of Oxford, 2000.

138. Peri, S. *et al.*, "Human protein reference database as a discovery resource for proteomics", *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D497–501, 2004.

139. Platt, A. *et al.*, "Conditions under which genome-wide association studies will be positively misleading", *Genetics*, vol. 186, no. 3, pp. 1045–52, 2010.

140. Price, A. L. *et al.*, "Principal components analysis corrects for stratification in genome-wide association studies", *Nat Genet*, vol. 38, no. 8, pp. 904–9, 2006.

141. Price, A. L. *et al.*, "New approaches to population stratification in genome-wide association studies", *Nat Rev Genet*, vol. 11, no. 7, pp. 459–63, 2010.

142. Pritchard, J. K. *et al.*, "Inference of population structure using multilocus genotype data", *Genetics*, vol. 155, no. 2, pp. 945–59, 2000.

143. Purcell, S. *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses", *Am J Hum Genet*, vol. 81, no. 3, pp. 559–75, 2007.

144. Pusapati, R. V. *et al.*, "ATM promotes apoptosis and suppresses tumorigenesis in response to Myc", *PNAS*, vol. 103, no. 5, pp. 1446–1451, 2006.

145. Ramanan, V. K. *et al.*, "Pathway analysis of genomic data: concepts, methods, and prospects for future development", *Trends Genet*, vol. 28, no. 7, pp. 323–32, 2012.

146. Raychaudhuri, S. *et al.*, "Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions", *PLoS Genet*, vol. 5, no. 6, e1000534, 2009.

147. Rebbeck, T. R. *et al.*, "Hormone-dependent effects of FGFR2 and MAP3K1 in breast cancer susceptibility in a population-based sample of post-menopausal African-American and European-American women", *Carcinogenesis*, vol. 30, no. 2, pp. 269–74, 2009.

148. Rioux, J. D. *et al.*, "Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis", *Nat Genet*, vol. 39, no. 5, pp. 596–604, 2007.

149. Risch, N. *et al.*, "The future of genetic studies of complex human diseases", *Science*, vol. 273, no. 5281, pp. 1516–7, 1996.

150. Ritchie, M. D. *et al.*, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer", *Am J Hum Genet*, vol. 69, no. 1, pp. 138–47, 2001.

151. Rodriguez, J. A. *et al.*, "Functional centrality in graphs", *Linear and Multilinear Algebra*, vol. 55, no. 3, pp. 293–302, 2007.

152. Rohl, C. *et al.*, "Cataloging the relationships between proteins: A review of interaction databases", *Mol Biotech*, vol. 34, pp. 69–93, 2006.

153. Rosset, S. *et al.*, "The population genetics of chronic kidney disease: insights from the MYH9-APOL1 locus", *Nat Rev Nephrol*, vol. 7, no. 6, pp. 313–26, 2011.

154. Saxena, R. *et al.*, "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels", *Science*, vol. 316, no. 5829, pp. 1331–6, 2007.

155. Scardoni, G. *et al.*, "Analyzing biological network parameters with CentiScaPe", *Bioinformatics*, vol. 25, no. 21, pp. 2857–9, 2009.

156. Schadt, E. E., "Molecular networks as sensors and drivers of common human diseases", *Nature*, vol. 461, no. 7261, pp. 218–223, 2009.

157. Schaid, D. J. *et al.*, "Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease", *Am J Hum Genet*, vol. 63, no. 5, pp. 1492–506, 1998.

158. Segre, D. *et al.*, "Modular epistasis in yeast metabolism", *Nat Genet*, vol. 37, no. 1, pp. 77–83, 2005.

159. Shahbaba, B. *et al.*, "A pathway analysis method for genome-wide association studies", *Stat Med*, vol. 31, no. 10, pp. 988–1000, 2012.

160. Sharma, A. *et al.*, "Network-based Analysis of Genome Wide Association Data Provides Novel Candidate Genes for Lipid and Lipoprotein Traits", *Mol Cell Proteomics*, vol. 12, no. 11, pp. 3398–408, 2013.

161. Shifman, S., "Linkage disequilibrium patterns of the human genome across populations", *Hum Mol Genet*, vol. 12, no. 7, pp. 771–776, 2003.

162. Shirokov, Y. M., "Algebra of one-dimensional generalized functions", *Theor Math Phys*, vol. 39, no. 3, pp. 471–477, 1979.

163. Siegel, S. *et al.*, *Non-parametric statistics for the behavioral sciences*, vol. 2, McGraw-Hill Humanities, 1998.

164. Sladek, R. *et al.*, "A genome-wide association study identifies novel risk loci for type 2 diabetes", *Nature*, vol. 445, no. 7130, pp. 881–5, 2007.

165. Slatkin, M., *The age of alleles. In Modern Developments in Theoretical Population Genetics, 3rd ed*, Oxford Univ. Press. Oxford Univ. Press., 2002, pp. 233–258.

166. Spencer, C. C. *et al.*, "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip", *PLoS Genet*, vol. 5, no. 5, e1000477, 2009.

167. Stark, C. *et al.*, "BioGRID: a general repository for interaction datasets", *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D535–9, 2006.

168. Stranger, B. E. *et al.*, "Progress and promise of genome-wide association studies for human complex trait genetics", *Genetics*, vol. 187, no. 2, pp. 367–83, 2011.

169. Strangio, M. A., *Graph-based Exploratory Analysis of Biological Interaction Networks*, Adv Tech, Kankesu Jay (Ed.), 2009.

170. Subramanian, A. *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles", *Proc Natl Acad Sci U S A*, vol. 102, no. 43, pp. 15545–50, 2005.

171. Szumilas, M., "Explaining odds ratios", *J Can Acad Child Adolesc Psychiatry*, vol. 19, no. 3, pp. 227–9, 2010.

172. Tevfik, D. M., *Introduction to genetic epidemiology*, http://www.dorak.info/epi/genetepi.html, [Online; accessed March2014], 2009.

173. Thomas, D. C. *et al.*, "Point: population stratification: a problem for case-control studies of candidate-gene associations?", *Cancer Epidemiol Biomarkers Prev*, vol. 11, no. 6, pp. 505–12, 2002.

174. Thye, T. *et al.*, "Common variants at 11p13 are associated with susceptibility to tuberculosis", *Nat Genet*, vol. 44, no. 3, pp. 257–9, 2012.

175. Tintle, N. L. *et al.*, "Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16", *BMC Proc*, vol. 3 Suppl 7, S96, 2009.

176. Torkamani, A. *et al.*, "Pathway analysis of seven common diseases assessed by genome-wide association", *Genomics*, vol. 92, no. 5, pp. 265–72, 2008.

177. Visscher, P. M., "Sizing up human height variation", *Nat Genet*, vol. 40, no. 5, pp. 489–90, 2008.

178. Visscher, P. M., "Handbook on Analyzing Human Genetic Data: Computational Approaches and Software edited by LIN, S. and ZHAO, H", *Biometrics*, vol. 66, no. 4, pp. 1310–1310, 2010.

179. Visscher, P. M. *et al.*, "Five years of GWAS discovery", *Am J Hum Genet*, vol. 90, no. 1, pp. 7–24, 2012.

180. Voight, B. F. *et al.*, "Confounding from cryptic relatedness in case-control association studies", *PLoS Genet*, vol. 1, no. 3, e32, 2005.

181. Wang, K. *et al.*, "Pathway-based approaches for analysis of genomewide association studies", *Am J Hum Genet*, vol. 81, no. 6, pp. 1278–83, 2007.

182. Wang, K. *et al.*, "Analysing biological pathways in genome-wide association studies", *Nat Rev Genet*, vol. 11, no. 12, pp. 843–54, 2010.

183. Wasserman, L., *All of statistics*, Springer New Yorg, 2004.

184. Wei, C. L. *et al.*, "A global map of p53 transcription-factor binding sites in the human genome", *Cell*, vol. 124, no. 1, pp. 207–19, 2006.

185. Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.", *Nat Rev Genet*, vol. 447, no. 1, pp. 661–678, 2007.

186. Wilcoxon, F., "Individual comparisons by ranking methods", *Biomet Bul*, vol. 1, no. 6, pp. 80–83, 1945.

187. Willer, C. J. *et al.*, "METAL: fast and efficient meta-analysis of genomewide association scans", *Bioinformatics*, vol. 26, no. 17, pp. 2190–1, 2010.

188. Wills, C., "Principles of Population Genetics, 4th edition", *Journal of Heredity*, vol. 98, no. 4, pp. 382–382, 2007.

189. Wright, F A. *et al.*, "Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2", *Nat Genet*, vol. 43, no. 6, 539âĂŞ546, 2011.

190. Wu, J. *et al.*, "Integrated network analysis platform for protein-protein interactions", *Nat Methods*, vol. 6, no. 1, pp. 75–7, 2009.

191. Wu, M. C. *et al.*, "Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways", *Stat Methods Med Res*, vol. 18, no. 6, pp. 577–93, 2009.

192. Xenarios, I., "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions", *Nuc Ac Res*, vol. 30, no. 1, pp. 303–305, 2002.

193. Yang, W. *et al.*, "Variable set enrichment analysis in genome-wide association studies", *Eur J Hum Genet*, vol. 19, no. 8, pp. 893–900, 2011.

194. Yildirim, M. A. *et al.*, "Drug-target network", *Nat Biotechnol*, vol. 25, no. 10, pp. 1119–26, 2007.

195. Yu, J. *et al.*, "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness", *Nat Genet*, vol. 38, no. 2, pp. 203–8, 2006.

196. Yu, K. *et al.*, "Pathway analysis by adaptive combination of P-values", *Genet Epidemiol*, vol. 33, no. 8, pp. 700–9, 2009.

197. Zaykin, D. V. *et al.*, "Truncated product method for combining P-values", *Genet Epidemiol*, vol. 22, no. 2, pp. 170–85, 2002.

198. Zaykin, D. V. *et al.*, "Ranks of genuine associations in whole-genome scans.", *Genetics*, vol. 171, no. 1, pp. 813–823, 2005.

199. Zhang, K. *et al.*, "i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study", *Nucleic Acids Res*, vol. 38, no. Web Server issue, W90–5, 2010.

200. Zhang, Z. *et al.*, "Mixed linear model approach adapted for genome-wide association studies", *Nat Genet*, vol. 42, no. 4, pp. 355–60, 2010.

201. Zheng, G. *et al.*, *Analysis of Genetic Association Studies*, Statistics for Biology and Health, Springer US, 2012.

202. Zhou, X. *et al.*, "Genome-wide efficient mixed-model analysis for association studies", *Nat Genet*, vol. 44, no. 7, pp. 821–4, 2012.

203. Zhu, X. *et al.*, "Getting connected: analysis and principles of biological networks", *Genes Dev*, vol. 21, no. 9, pp. 1010–24, 2007.

APPENDIX A

Supplementary materials

Table A.1: Module genes of ancGWAS on a simulated disease in an admixed population. (Subsection 3.2.2). Complete list of genes of modules using ancGWAS where each module in the main text is represent by its hub.

Module hub	Genes in module
TRAF3IP3	*CCT8*, *FGFR1OP2*, *SIKE1*, *STK24*, *TRAF3IP3*, *CTTNBP2NL*, *FAM40A*
CTTNBP2NL	*FGFR1OP2*, *SIKE1*, *STK24*, *STRN*, *TRAF3IP3*, *CTTNBP2NL*, *FAM40A*
RPSA	*SI*, *SLC2A5*, *RPSA*, *CCL5*
SLC2A5	*SLC2A5*, *SI*, *RPSA*, *CCL5*
SRSF10	*MARK1*, *ZC3H11A*, *PI4KB*, *RORC*, *RPL31*, *SFN*, *RBM34*, *PKP3*, *SRSF10*, *YWHAB*, *C1orf74*, *KIAA0101*, *POU5F1*, *GJA5*
LEPR	*LEPROT*, *PTPN11*, *TIE1*, *LEPR*
IPO13	*HDAC1*, *HSPA6*, *IPO13*, *RPL7*, *CRX*, *RPL5*, *PSMD2*, *CBX5*, *RPS8*, *NEDD8*, *CRABP2*, *HNRNPR*, *FBXO42*, *PPM1J*, *RAD21*, *LGALS3BP*, *HIST3H2BB*, *RPL11*, *RPL12*, *IPO7*, *UBE2I*, *WDTC1*, *TP73*, *RPL10A*, *HSP90AA1*, *ACTB*, *YBX1*
SNRPE	*HDAC1*, *UBE2I*, *ETV3*, *DDX20*, *CRABP2*, *SLX1A*, *NR0B2*, *HNRNPA1*, *PPM1J*, *TP73*, *GEMIN5*, *LRRC41*, *DCUN1D1*, *STXBP3*, *RGSL1*, *PPARD*, *PEF1*, *SNRPE*
CTSD	*PSMD4*, *PSMD2*, *ANP32E*, *PKP1*, *KRT8*, *DCLRE1B*, *S100A14*, *FLG*, *S100A16*, *FKBP8*, *UBQLN4*, *ATG5*, *PSMC1*, *LYST*, *CUL2*, *CENPJ*, *PSMA5*, *CTSD*, *TXN2*, *BGLAP*, *USP14*, *MGMT*, *UCHL5*, *CTSS*, *SLPI*, *SHFM1*, *PSMB4*, *GSTM2*, *GSTM3*, *PSMD10*, *PSAT1*, *ATP6V0A1*, *WFS1*

Continued on next page

Table A.1 – continued from previous page

Module hub	Genes in module
PSMA5	*PSMD4*, *PSMD2*, *ANP32E*, *PKP1*, *KRT8*, *DCLRE1B*, *S100A14*, *FLG*, *S100A16*, *FKBP8*, *UBQLN4*, *PSMA5*, *PSMC1*, *LYST*, *CUL2*, *ATG5*, *CTSD*, *TXN2*, *BGLAP*, *USP14*, *MGMT*, *UCHL5*, *SHFM1*, *PSMB4*, *GSTM2*, *GSTM3*, *PSMD10*, *PSAT1*, *ATP6V0A1*, *WFS1*
S100A14	*PSMD4*, *PSMD2*, *ANP32E*, *PKP1*, *KRT8*, *DCLRE1B*, *S100A14*, *FLG*, *S100A16*, *FKBP8*, *UBQLN4*, *PSMA5*, *PSMC1*, *LYST*, *CUL2*, *ATG5*, *CTSD*, *TXN2*, *BGLAP*, *USP14*, *MGMT*, *UCHL5*, *SHFM1*, *PSMB4*, *GSTM2*, *GSTM3*, *PSMD10*, *PSAT1*, *ATP6V0A1*, *WFS1*
PSMB4	*PSMD4*, *PSMD2*, *ANP32E*, *PKP1*, *KRT8*, *DCLRE1B*, *S100A14*, *FLG*, *S100A16*, *FKBP8*, *UBQLN4*, *PSMA5*, *PSMC1*, *LYST*, *CUL2*, *ATG5*, *CTSD*, *TXN2*, *BGLAP*, *USP14*, *MGMT*, *UCHL5*, *SHFM1*, *PSMB4*, *GSTM2*, *GSTM3*, *PSMD10*, *PSAT1*, *ATP6V0A1*, *WFS1*
RPL31	*SNIP1*, *OLFML3*, *PPP1R8*, *SRSF10*, *S100A10*, *KIAA0101*, *CRABP2*, *RORC*, *PPM1J*, *RPL22*, *RBM34*, *PRKRA*, *SETDB1*, *POU5F1*, *C1orf74*, *GIPC2*, *TMBIM4*, *TARDBP*, *UBE2I*, *HDGF*, *ZC3H11A*, *TP73*, *RPL31*, *SFN*, *CDC42*
CGN	*EXOSC10*, *TJP1*, *STAT3*, *RBM8A*, *DIRAS3*, *CGN*, *GJA8*, *RRAS2*, *ARHGEF2*, *YWHAH*, *ISG15*, *PPP6R3*, *STMN1*, *IL22RA1*, *USP21*, *IL23R*
STXBP3	*HDAC1*, *UBE2I*, *ETV3*, *DDX20*, *CRABP2*, *SLX1A*, *NR0B2*, *HNRNPA1*, *PPM1J*, *TP73*, *GEMIN5*, *DCUN1D1*, *STXBP3*, *PPARD*, *PEF1*, *SNRPE*
NID1	*FCGR3A*, *LAX1*, *LACRT*, *ADAM15*, *ZC3H12A*, *SHC1*, *PTPN22*, *TANK*, *CRABP2*, *PPM1J*, *NID1*, *PTPRD*, *CD48*, *KHDRBS1*, *MUC1*, *SH2D2A*, *CSF3R*, *LAMC1*, *UBE2I*, *TP73*, *PTPRF*, *STAT5A*, *CD247*, *HSP90AA1*, *LCK*, *CDC42*
RPAP2	*HSPA6*, *RPL7*, *HSPA8*, *PABPC4*, *KHDRBS3*, *PSMD2*, *TP73*, *PITX2*, *RPS8*, *NEDD8*, *CRABP2*, *HNRNPR*, *PPM1J*, *JAK1*, *EIF2C4*, *RPL5*, *INTS7*, *INTS6*, *FBXO42*, *GBP2*, *CREB1*, *LGALS3BP*, *EIF2C3*, *RPL12*, *UBE2I*, *RPS6KA1*, *WDTC1*, *RPL35*, *RPAP2*, *TUBB*, *MYCBP*, *RPL10A*, *HSP90AA1*, *ACTB*, *YBX1*
SFN	*ZC3H11A*, *PI4KB*, *EXO1*, *RORC*, *RPL31*, *SFN*, *RBM34*, *PKP3*, *MARK1*, *YWHAB*, *C1orf74*, *KIAA0101*, *RIMS3*, *SRSF10*, *TESK2*, *GJA5*
LAMC1	*UBE2I*, *TANK*, *CRABP2*, *PPM1J*, *TP73*, *PTPRF*, *NID1*, *LACRT*, *ZC3H12A*, *LAMC1*, *LCK*, *PTPRD*
APOA2	*APOA1BP*, *APOD*, *APOA2*

Table A.2: 129 previously confirmed breast cancer susceptibility genes.

Gene	Chr	Journal Reference	Gene	Chr	Journal Reference
GSTM1	1	Helzlsouer, K.J. et *al*. 1998. PMID:9539246	*GSTP1*	11	Nordgard, S. H. et *al*. 2007. PMID:17301692
GSTM3	1	Mitrunen, K. et *al*. 2001. PMID:11303592	*HRAS*	11	Nefedov MD et *al*. 1990. PMID:2086347
Intergenic	1	Thomas et *al*. 2009. PMID:19330030	*LSP1*	11	Easton et *al*. 2007. PMID:17529967
LEPR	1	Gallicchio, L. et *al*. 2007. PMID:17428620	*MMP1*	11	Przybylowska, K. et *al*. 2004. PMID:15149160
MTHFR	1	Martha, J. S. et *al*. 2005. PMID:15868433	*PGR*	11	Pooley, K. A. et *al*. 2006. PMID:16614108
MYCL1	1	Champeme MH et *al*. 1992. PMID:1345822	*SIPA1*	11	Crawford, N. P. et *al*. 2006. PMID:16563182
PTGS2	1	Langsenlehner, U. et *al*. 2006. PMID:16489098	*CDKN1B*	12	Ma, H. et *al*. 2006. PMID:16804901
TNFRSF1B	1	Mestiri, S. et *al*. 2005. PMID:15863392	*IFNG*	12	Kamali-Sarvestani, E. et *al*. 2005. PMID:15890243
TP73	1	Li, H. et *al*. 2006. PMID:16950799	*IGF1*	12	Al-Zahrani, A. et *al*. 2006. PMID:16306136
BARD1	2	Huo, X. et *al*. 2006. PMID:17028982	*Intergenic*	12	Murabito et *al*. 2007. PMID:17903305
CASP8	2	Cox, A. et al et *al*. 2007. PMID:17293864	*LRP1*	12	Benes, P. et *al*. 2003. PMID:12793904
CTLA4	2	Ghaderi, A. et *al*. 2004. PMID:15218356	*MDM2*	12	Wasielewski, M. et *al*. 2006. PMID:17080308
CYP1B1	2	Matyjasik, J. et *al*. 2007. PMID:17458695	*VDR*	12	Lundin AC et *al*. 1999. PMID:10344739
Intergenic	2	Stacey et *al*. 2007. PMID:17529974	*ABCC4*	13	Murabito et *al*. 2007. PMID:17903305
LHCGR	2	Powell, B. L. et *al*. 2003. PMID:12679452	*BRCA2*	13	Golshan, M. et *al*. 2006. PMID:16769276
SRD5A2	2	Yang, C. et *al*. 2002. PMID:12100746	*LIG4*	13	Bau, D. T. et *al*. 2004. PMID:15256476
UGT1A1	2	Mol Biol (Mosk). 2006. 40(2). 263-70. PMID:16637266	*RB1*	13	Berns EM et *al*. 1995. PMID:7615356

Continued on next page

Table A.2 – continued from previous page

Gene	Chr	Journal Reference	Gene	Chr	Journal Reference
ALDH1L1	3	Stevens, V. L. et *al.* 2007. PMID:17548676	*ESR2*	14	Gallicchio, L. et *al.* 2006. PMID:16808847
GPX1	3	Cox, D. G. et *al.* 2006. PMID:16945136	*MTHFD1*	14	Stevens, V. L. et *al.* 2007. PMID:17548676
PIK3CA	3	Breast cancer research and treatment. 2005. PMID:16317585	*RAD51L1*	14	Thomas et *al.* 2009. PMID:19330030
RASSF1	3	Schagdarsurengin, U. et *al.* 2005. PMID:15942659	*XRCC3*	14	Bewick, M. A. et *al.* 2006. PMID:17116943
PPARGC1A	4	Wirtenberger, M. et *al.* 2006. PMID:16704985	*AKAP13*	15	Wirtenberger, M. et *al.* 2005. PMID:16234258
SULT1E1	4	Choi, J. Y. et *al.* 2005. PMID:15894657	*CYP11A1*	15	Zheng, W. et *al.* 2004. PMID:15159300
ADRB2	5	Huang XE et *al.* 2001. PMID:11434877	*CYP19A1*	15	Miyoshi Y et *al.* 2000. PMID:10956405
FGFR4	5	Thussbas, C. et *al.* 2006. PMID:16822847	*CYP1A1*	15	Long, J. R. et *al.* 2007. PMID:17429315
Intergenic	5	Easton et *al.* 2007. PMID:17529967	*CYP1A2*	15	Kotsopoulos, J. et *al.* 2007. PMID:17507615
MAP3K1	5	Easton et *al.* 2007. PMID:17529967	*FBN1*	15	Murabito et *al.* 2007. PMID:17903305
PPARGC1B	5	Wirtenberger, M. et *al.* 2006. PMID:16704985	*RAD51*	15	Jakubowska, A. et *al.* 2003. PMID:12750242
PRLR	5	Vaclavicek, A. et *al.* 2006. PMID:16434456	*ERCC4*	16	Smith, T. R. et *al.* 2003. PMID:14652281
XRCC4	5	Allen-Brady, K. et *al.* 2006. PMID:16835328	*GLG1*	16	Kibriya et *al.* 2008. PMID:18463975
C6orf97	6	Zheng et *al.* 2009. PMID:19219042	*MMP2*	16	Zhou Y 2004. PMID:14604886
CDKN1A	6	Staalesen, V. et *al.* 2006. PMID:17062672	*NQO1*	16	Menzel, H. J. et *al.* 2004. PMID:15138483
ECHDC1	6	Gold et *al.* 2008. PMID:18326623	*SULT1A1*	16	Mol Biol (Mosk). 2006. 40(2). 263-70. PMID:16637266
ESR1	6	Gallicchio, L. et *al.* 2006. PMID:16808847	*TOX3*	16	Thomas et *al.* 2009. PMID:19330030
GSTA1	6	Sweeney, C. et *al.* 2003. PMID:12516103	*CASC16*	16	Easton et *al.* 2007. PMID:17529967

Continued on next page

Table A.2 – continued from previous page

Gene	Chr	Journal Reference	Gene	Chr	Journal Reference
HLA-DQB1	6	Chaudhuri, S. et *al*. 2000. PMID:11027344	*TNRC9*	16	Stacey et *al*. 2007. PMID:17529974
HLA-DRB1	6	Chaudhuri, S. et *al*. 2000. PMID:11027344	*ACACA*	17	Sinilnikova, O. M. et *al*. 2004. PMID:15333468
PRL	6	Vaclavicek, A. et *al*. 2006. PMID:16434456	*AKAP10*	17	Wirtenberger, M. et *al*. 2006. PMID:16956908
RNF146	6	Gold et *al*. 2008. PMID:18326623	*BRCA1*	17	Fu, X. et *al*. 2007. PMID:17557253
SOD2	6	Yao S et *al*. 2010. PMID:20309628	*COL1A1*	17	Murabito et *al*. 2007. PMID:17903305
VEGFA	6	Krippl, P. et *al*. 2003. PMID:12845639	*ERBB2*	17	Tommasi, S. et *al*. 2007. PMID:17452776
ABCB1	7	Zubor, P. et *al*. 2007. PMID:17549370	*GH1*	17	Endocrine-related cancer. 2005 Dec:12(4):917-28. PMID:16322331
AHR	7	Pharmacogenet Genomics. 2006. 16(4). 237-43. PMID:16538170	*HER2*	17	Ameyaw MM et *al*. 2002. PMID:12166652
CYP3A5	7	Tucker, A. N. et *al*. 2005. PMID:15596297	*HSD17B1*	17	Wu AH 2003. PMID:12584742
IGFBP3	7	Al-Zahrani, A. et *al*. 2006. PMID:16306136	*ITGB3*	17	Wang-Gohrke, S. et *al*. 2004. PMID:15609125
IL6	7	Snoussi, K. et *al*. 2005. PMID:16464738	*MPO*	17	Ahn, J. et *al*. 2004. PMID:15492293
Intergenic	7	Murabito et *al*. 2007. PMID:17903305	*SHBG*	17	Cui, Y. et *al*. 2005. PMID:15894658
NOS3	7	Lee, K. M. et *al*. 2007. PMID:17262178	*TIMP2*	17	Zhou Y 2004. PMID:14604886
PON1	7	Gallicchio, L. et *al*. 2007. PMID:17428620	*TP53*	17	Thangarajan Rajkumar et *al*. 2008. PMID:18058229
POR	7	Haiman, C. A. et *al*. 2007. PMID:17440066	*Intergenic*	18	Murabito et *al*. 2007. PMID:17903305

Continued on next page

Table A.2 – continued from previous page

Gene	Chr	Journal Reference	Gene	Chr	Journal Reference
SERPINE1	7	Lei, H. et *al*. 2007. PMID:17616807	*GPX4*	19	Nasim Mavaddat , et *al*. Cancer epidemiology, biomarkers & prevention 2009 18(1):255-9. PMID:19124506
XRCC2	7	Kuschel, B. et *al*. 2002. PMID:12023982	*KLK3*	19	Yang, Q. et *al*. 2002. PMID:12168876
Intergenic	8	Easton et *al*. 2007. PMID:17529967	*TGFB1*	19	Thangarajan Rajkumar et *al*. 2008. PMID:18058229
MYC	8	Wirtenberger, M. et *al*. 2005. PMID:15929079	*UQCRFS1*	19	Ohashi Y 2004. PMID:15047214
NAT1	8	Pfau W 1998. PMID:9829711	*XRCC1*	19	Bewick, M. A. et *al*. 2006. PMID:17116943
NAT2	8	Christine B Ambrosone et *al*. 2008. PMID:18187392	*AURKA*	20	Lo, Y. L. et *al*. 2005. PMID:15688402
NBN	8	Lu, J. et *al*. 2006. PMID:16714331	*GNAS*	20	Otterbach, F. et *al*. 2006. PMID:17186357
POLB	8	Sliwinski, T. et *al*. 2006. PMID:17131038	*MMP9*	20	Grieu, F. et *al*. 2004. PMID:15609121
TNFRSF10A	8	Frank, B. et *a*l. 2005. PMID:15975957	*CBS*	21	Stevens, V. L. et *a*l. 2007. PMID:17548676
PTCH	9	Chang-Claude J et *al*. 2003. PMID:12516098	*COL18A1*	21	Balasubramanian, S. P. et *al*. 2007. PMID:17587451
CYP17	10	Wu AH 2003. PMID:12584742	*CHEK2*	22	Meyer, A. et *al*. 2007. PMID:17250914
CYP17A1	10	Piller, R. et *al*. 2006. PMID:16702327	*COMT*	22	Song, C. G. et *al*. 2006. PMID:17217814
CYP2C19	10	Werner Schroth et *al*. 2007. PMID:18024866	*CYP2D6*	22	Werner Schroth et *a*l. 2007. PMID:18024866
FGFR2	10	Hunter et *al*. 2007. PMID:17529973	*EP300*	22	Wirtenberger, M. et *al*. 2006. PMID:16704985
MGMT	10	Han, J. et *al*. 2006. PMID:16788379	*GSTT1*	22	Mitrunen, K. et *al*. 2001. PMID:11303592
ATM	11	Sommer, S. S. et *al*. 2002. PMID:11996792	*AR*	X	Yu H et *al*. 2000. PMID:10817350

Table A.3: Top 20 sub-networks, and related pathway enrichment results from dmGWAS. Modules are ranked by their z-score, ZM, from dmGWAS. For each pathway, the z-score (ZP) is reported representing the level of its similarity with the related module, which is also a function of the size of the module. Different scores of overlaps between the module, the pathway and the set of known genes are also reported.

Hub	ZM	MS	Pathway	ZP	MOG	MKG	PKG
COPS6	10.54	8	PI3K/AKT activation	147.7851	1	1	0
BCAR3	10.52	7	Proteoglycan syndecan-mediated signaling events	1.8236	2	1	2
SPTLC1	10.44	8	Proteoglycan syndecan-mediated signaling events	1.2892	1	1	2
ETS1	10.30	7	Proteoglycan syndecan-mediated signaling events	3.9929	2	1	2
DMWD	10.21	7	Proteoglycan syndecan-mediated signaling events	2.683	2	1	2
GALK1	10.18	8	Proteoglycan syndecan-mediated signaling events	3.7358	3	1	2
CBL	10.14	5	Proteoglycan syndecan-mediated signaling events	1.8236	2	1	2
ATP6V1E1	10.13	8	Hemostasis	35.3949	3	1	0
MLH1	10.12	6	Proteoglycan syndecan-mediated signaling events	2.8943	3	1	2
RPS14	10.08	7	TRAIL signaling pathway	2.7755	2	1	2
RORA	10.06	6	LKB1 signaling events	4.024	4	1	2
MDM2	10.04	6	Proteoglycan syndecan-mediated signaling events	2.683	2	1	2
DSG1	10.04	8	Signaling events mediated by VEGFR1 and VEGFR2	3.123	3	1	2
RPS27A	10.03	7	Cell Cycle, Mitotic	9.8726	1	1	1
PTN	9.99	5	FAS (CD95) signaling pathway	42.752	2	1	1
ATXN7	9.97	6	Metabolism of proteins	28.0057	2	1	0
UBB	9.93	6	Proteoglycan syndecan-mediated signaling events	4.8908	3	1	2
LRP2	9.92	6	Hemostasis	27.0515	2	1	0
PIAS2	9.90	7	Proteoglycan syndecan-mediated signaling events	2.2339	3	1	2
HNF4A	9.86	7	Proteoglycan syndecan-mediated signaling events	1.2892	1	1	2

Abbreviation. P: P-value; $AdjP$: Adjusted p-value; MS: Module size; ZP: Pathway z-score; MOG: Gene overlapping between the module and the pathway; MKG: Number of genes overlapping between the module and the set of arbitrarily chosen genes as known disease genes; PKG: Number of genes overlapping between the pathway and the set of arbitrarily chosen genes as known disease genes;

Table A.4: Module genes of ancGWAS on a case-control breast cancer data set (Hunter, D. J. *et al.*, 2007), from the CGEMS project. (Section 4.2). Complete list of genes from modules using ancGWAS, where each module in the main text is represented by its hub.

Module hub	Genes in module
MAP3K12	*LHCGR* , *MBIP*, *SH3RF1*, *CXCL9*, *RPL18A*, *MFAP5*, *FBN1*, *CHST9*, *SPRY2*, *MAP3K12*, *MAPK6*, *RGS1*, *RABEPK*, *RHOB*, *GNAI2*, *CANX*, *PF4*
CDKN1A	*TTLL5*, *MYCL1*, *CDC5L*, *C1orf123*, *TXN*, *MAX*, *CCND2*, *GCKR*, *MKRN1*, *GADD45G*, *TNIP2*, *POLD2*, *RABEPK*, *TK1*, *CDC45*, *BAD*, *CIZ1*, *RPL18*, *CCDC85B*, *NR1H2*, *HPD*
UNC93B1	*SEL1L*, *CCDC47*, *PIGN*, *SPNS1*, *GOLT1B*, *NPC1*, *TM9SF3*, *TM9SF2*, *UNC93B1*, *STT3B*, *STT3A*, *YIPF5*, *HM13*
EIF2B1	*IL6*, *FBN1*, *ZBTB16*, *CXCL9*, *DCD*, *PDIA4*, *ARHGDIA*, *EIF2B3*, *IL6R*, *DCC*, *SPRY2*, *MFAP5*, *PF4*, *MRPL4*, *Adra2a*, *RHOB*, *ADRA2C*, *EIF2S2*
SMC3	*MYCL1*, *RPLP1*, *FBXO2*, *ANP32E*, *KIFAP3*, *MXD3*, *EIF3H*, *WFDC5*, *SMC3*, *TRAF3IP1*, *ANXA1*, *BECN1*, *STAG2*, *STAG3*, *RABEPK*, *SLC25A4*, *MCM3*, *USP19*, *WRAP73*, *SYCP3*, *H2BFM*, *CCNB1*, *MAX*, *FEZ2*, *REC8*, *PHB*
LNX2	*RIPK1*, *HIP1*, *CRADD*, *OTUD7B*, *PEA15*, *BID*, *CFLAR*, *NOD1*, *RNF138*, *RLIM*, *MALT1*, *PRKCI*, *LNX2*, *SHISA2*, *PARP2*, *PLEC*, *NUMB*, *CD8A*, *DEDD*, *NOL3*, *BCAP31*, *CASP10*, *UBE2Z*, *MAP3K14*, *APAF1*, *MAPT*, *FASLG*
ITGA5	*LHCGR*, *AUP1*, *RABIF*, *ITGA5*, *CHST9*, *ANGPTL3*, *L1CAM*, *GIPC1*, *FLT4*, *GNAI2*, *CANX*
ATF7IP	*ZBTB6*, *CSPG4*, *SLC12A4*, *SENP3*, *PKD1*, *SVEP1*, *MCAM*, *ATF7IP*, *GLG1*, *COL4A2*, *RRBP1*, *NFKB2*, *SPTBN1*, *CADM1*
GRIN2B	*CTLA4*, *DLG3*, *DLG2*, *FGF2*, *CDH2*, *AP1M1*, *FYN*, *RABEPK*, *PRKCG*, *PARK2*, *GRIN2B*, *STAT5B*, *CAMK2N1*, *FGF8*, *FGF7*, *FGF6*, *FGFR4*, *CD86*, *FGF3*, *LYN*
RTN4	*IL6*, *COL4A3BP*, *ACTG1*, *TRAF6*, *WWP1*, *RTN4*, *LSP1*, *ATL1*, *IL6R*, *CNTNAP1*, *RAD51AP1*, *PALB2*, *MAPKAPK2*, *ZBTB16*
PRKCG	*PDLIM5*, *GJA3*, *IGSF21*, *ANXA7*, *EXOC5*, *EEF1A1*, *PARD6A*, *PARD6B*, *PRKCG*, *GRIN2B*, *SULT1E1*, *GRM5*, *GRIA4*, *RGS2*, *UNC119*, *NOXA1*, *GABRA4*, *PTPN1*, *STXBP1*
MAPKAP1	*RPTOR*, *PECAM1*, *PDGFRA*, *CAAP1*, *GZMB*, *GULP1*, *CDC42EP1*, *PTK2B*, *CD36*, *PTPRZ1*, *RPS6*, *ECI1*, *CIB1*, *SETMAR*, *EPS8*, *DAB2*, *ABCC4*, *FBLN2*, *XRCC4*
PACSIN3	*YWHAZ*, *TGM2*, *ASAP1*, *PACSIN3*, *SPDYA*, *PON1*, *SUMO4*, *RABEPK*, *WIPF1*, *HSPA1A*, *DNM1*, *WASL*, *UBAC1*, *TRPV4*, *SYNJ1*, *ADAM12*, *PACSIN1*, *RHOA*, *HIST1H1B*, *FBL*

Continued on next page

Table A.4 – continued from previous page

Module hub	Genes in module
LEF1	*EEF1A1*, *MYCL1*, *CYP11A1*, *DPYSL2*, *FHIT*, *CTLA4*, *CDX1*, *MAX*, *NOTCH1*, *SMAD2*, *CD86*, *LYN*, *STAT5B*, *IGSF21*, *UNC119*, *PITX2*, *CYP11B2*, *LEF1*, *ALX4*, *SULT1E1*, *KPNA1*
ACVR1	*PLEKHB1*, *TGFBR1*, *ACVR1*, *ZBTB16*, *ENG*, *PLEKHJ1*, *NUAK2*, *RRAS2*, *NRAS*, *IGSF1*, *NEK8*, *IL6R*, *IL6*, *USP39*, *CHN1*, *DCAF6*, *GDF5*, *RHOJ*, *PLEK*, *DUSP13*, *RASD2*
PCBD2	*GTF2E1*, *ZZZ3*, *PCBD2*, *HES4*, *PALB2*, *RAD51AP1*, *MED7*, *WFDC1*, *ASCC2*, *PBXIP1*, *SSX3*
WAS	*PACSIN1*, *RHOQ*, *FYN*, *RBBP5*, *ARPC4*, *ARHGAP1*, *KDM4A*, *PTPRB*, *KDM6B*, *WAS*, *ENAH*, *BTK*, *CDC42*
BSG	*RFXANK*, *CAND1*, *SULT1E1*, *EEF1A1*, *PON1*, *SLC16A4*, *PPP2R1B*, *SUMO4*, *VHL*, *USP50*, *IGSF21*, *UNC119*, *ATXN10*, *HGS*, *RANBP3*
ARL4D	*DNAJA1*, *UBR1*, *NEK3*, *PRLR*, *ARL6IP1*, *PRKCSH*, *SNRPN*, *EML4*, *CSH1;*, *ARL4D*, *TMEM230*, *ZAP70*